Consistency without consensus

Distributed systems theory

Distributed programming

“The art of solving the same problem that you can solve on a single computer using multiple computers.”

—*book.mixu.net/distsys*

Distributed programming

“Generally a bad idea. Avoid if possible.”

—*me*

```
>>> x = 1
>>> print x
1
```

```
$ curl -Ss -XPOST 'http://db:8080/vars/x' -d'{"value":1}'
HTTP 502 (Bad Gateway)
$ curl -Ss -XGET 'http://db:8080/vars/x'
HTTP 500 (Internal Server Error)
```

Primitives and patterns

1980s: RPC

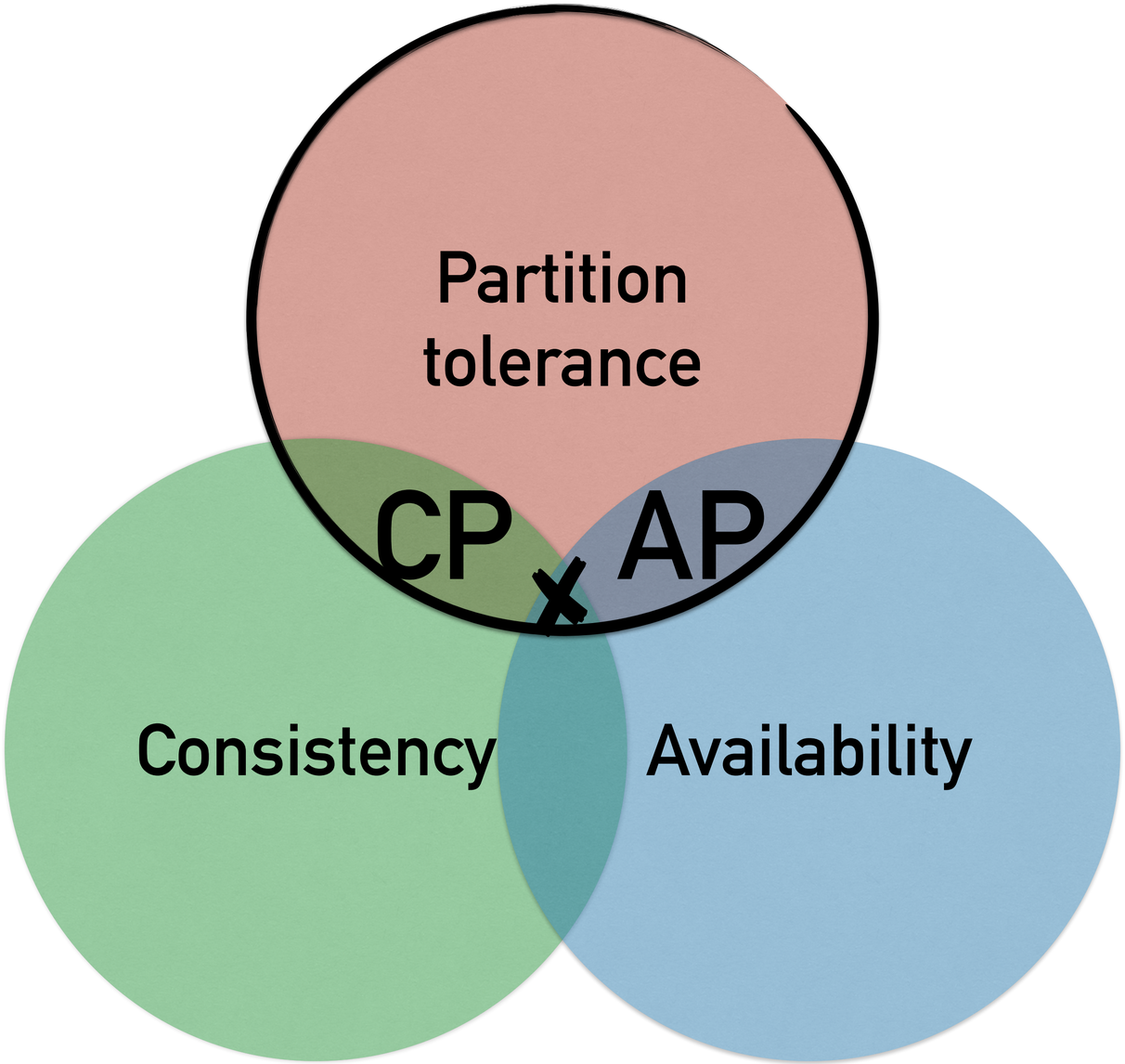
1990s: CORBA, middlewares

2000: CAP theorem

Partition-tolerance

“[That] the system continues to operate despite message loss due to network and/or node failure.”

—*Mixu*


Partition
tolerance
CP
AP
Consistency
Availability

CP

Paxos (doozer, chubby)

Zab (ZooKeeper)

Raft (Consul, etcd)

Viewstamped Replication (?)

AP

Cassandra

Riak

CouchBase

MongoDB

Failure

Delayed messages

Out-of-order messages

Dropped messages

Duplicate messages

CALM principle

Consistency

As

Logical

Monotonicity

ACID 2.0

Associative

Commutative

Idempotent

Distributed, sure, whatever

CRDT

Conflict-free

Replicated

Data

Type

Increment-only counter

Increment-only counter

"An increment-only counter is a replicated integer supporting operations **increment** to update, and **value** to query."

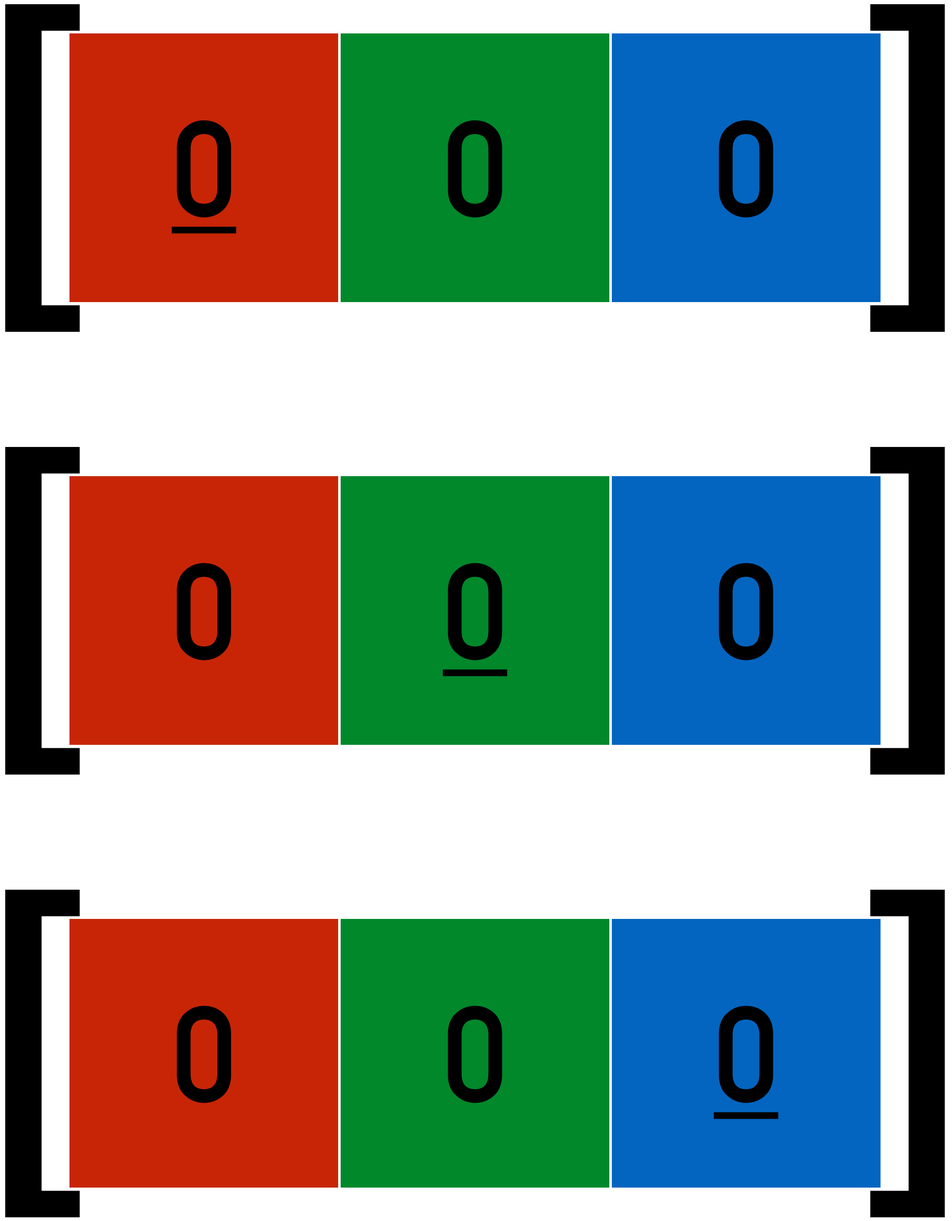
—*Shapiro*

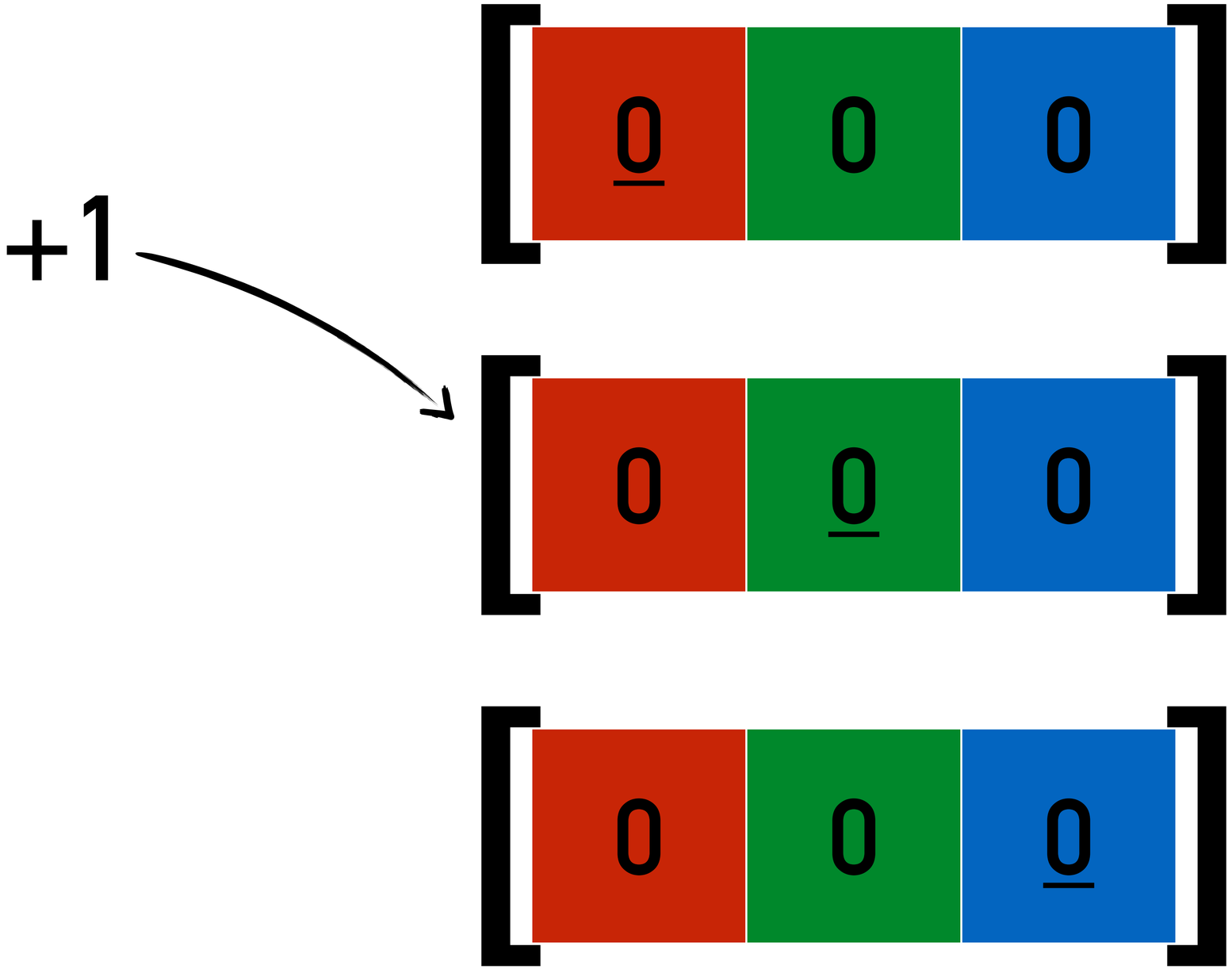
+

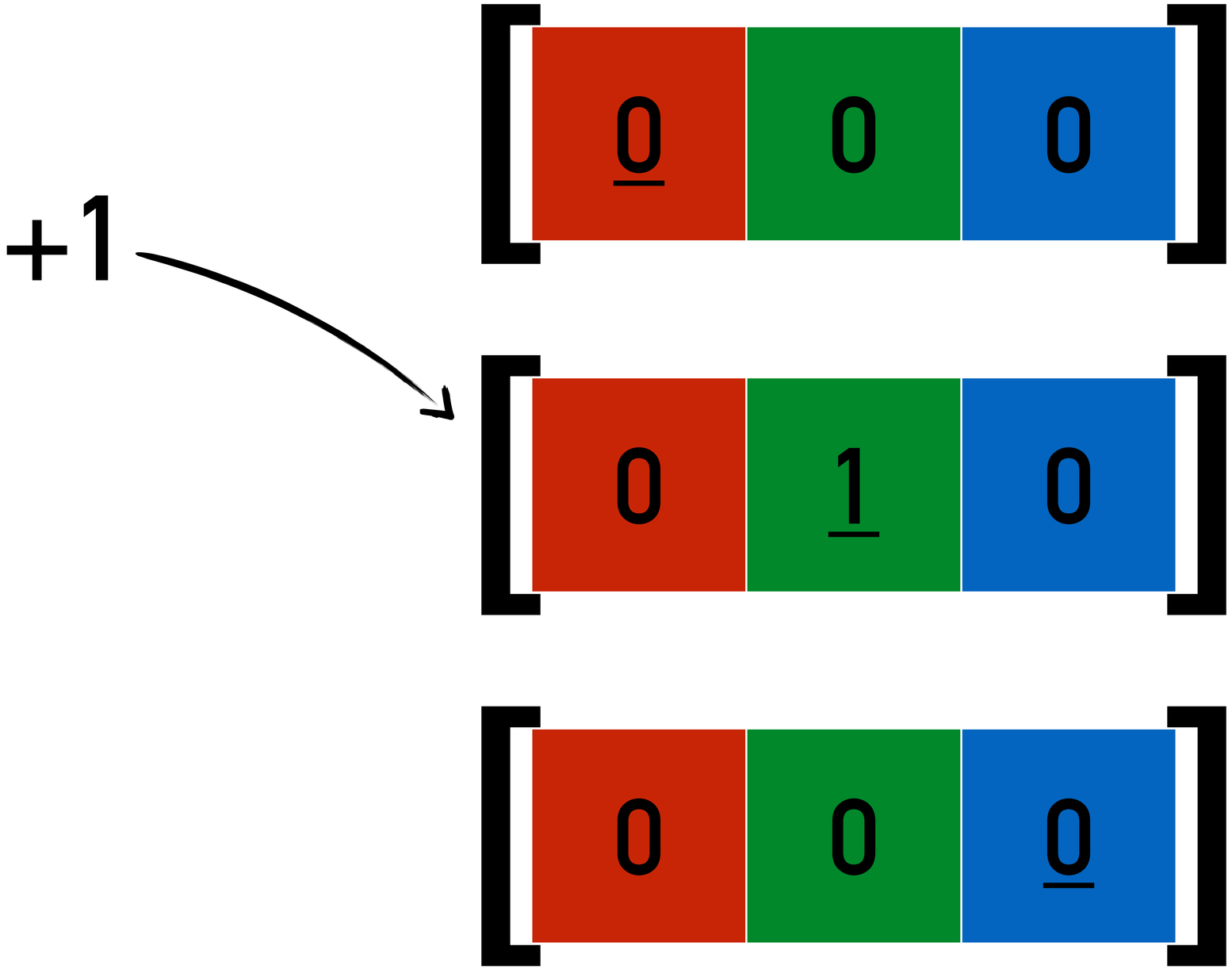
$$1 + (2 + 3) = (1 + 2) + 3 \quad ✔$$

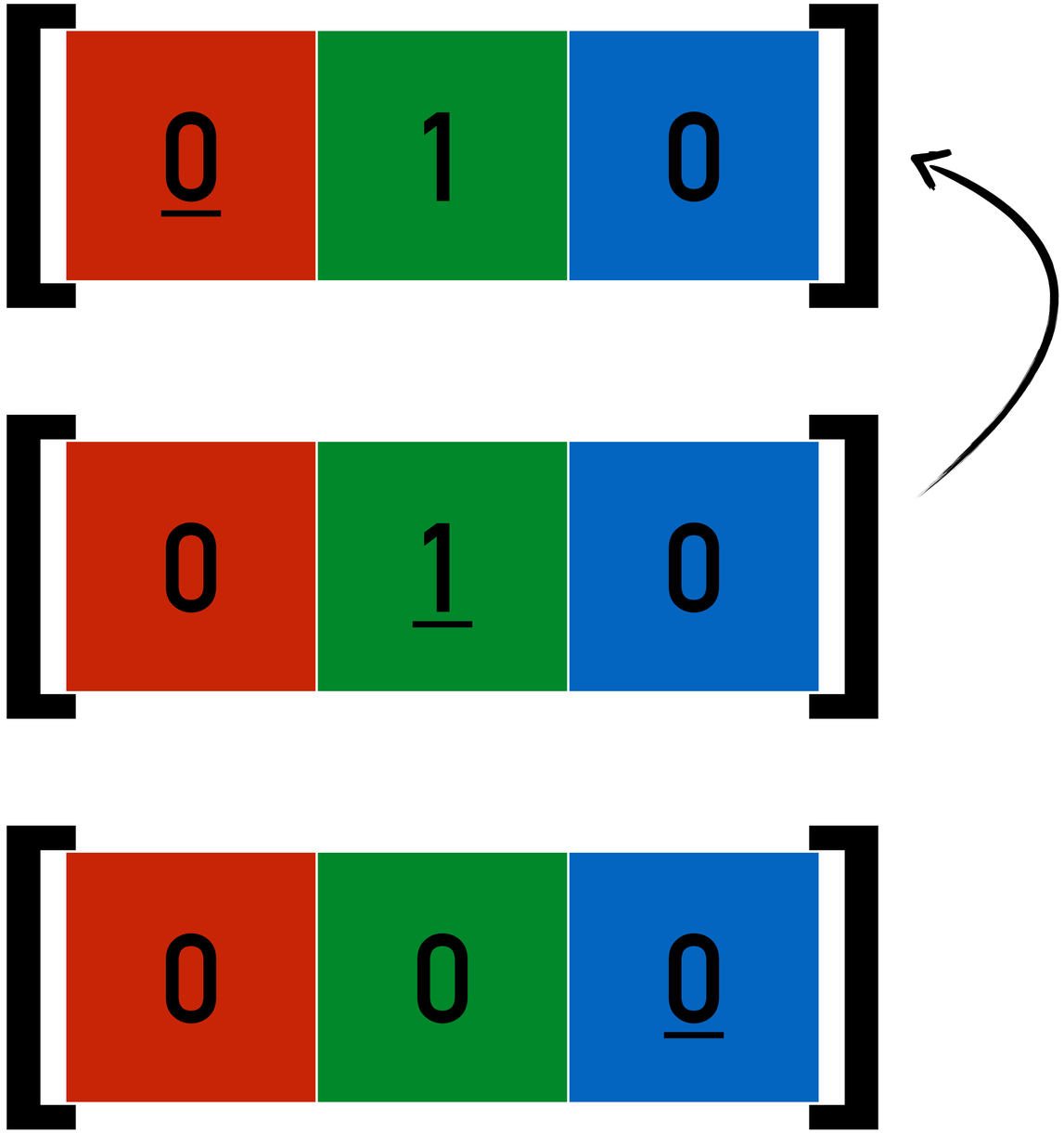
$$1 + 2 = 2 + 1 \quad ✔$$

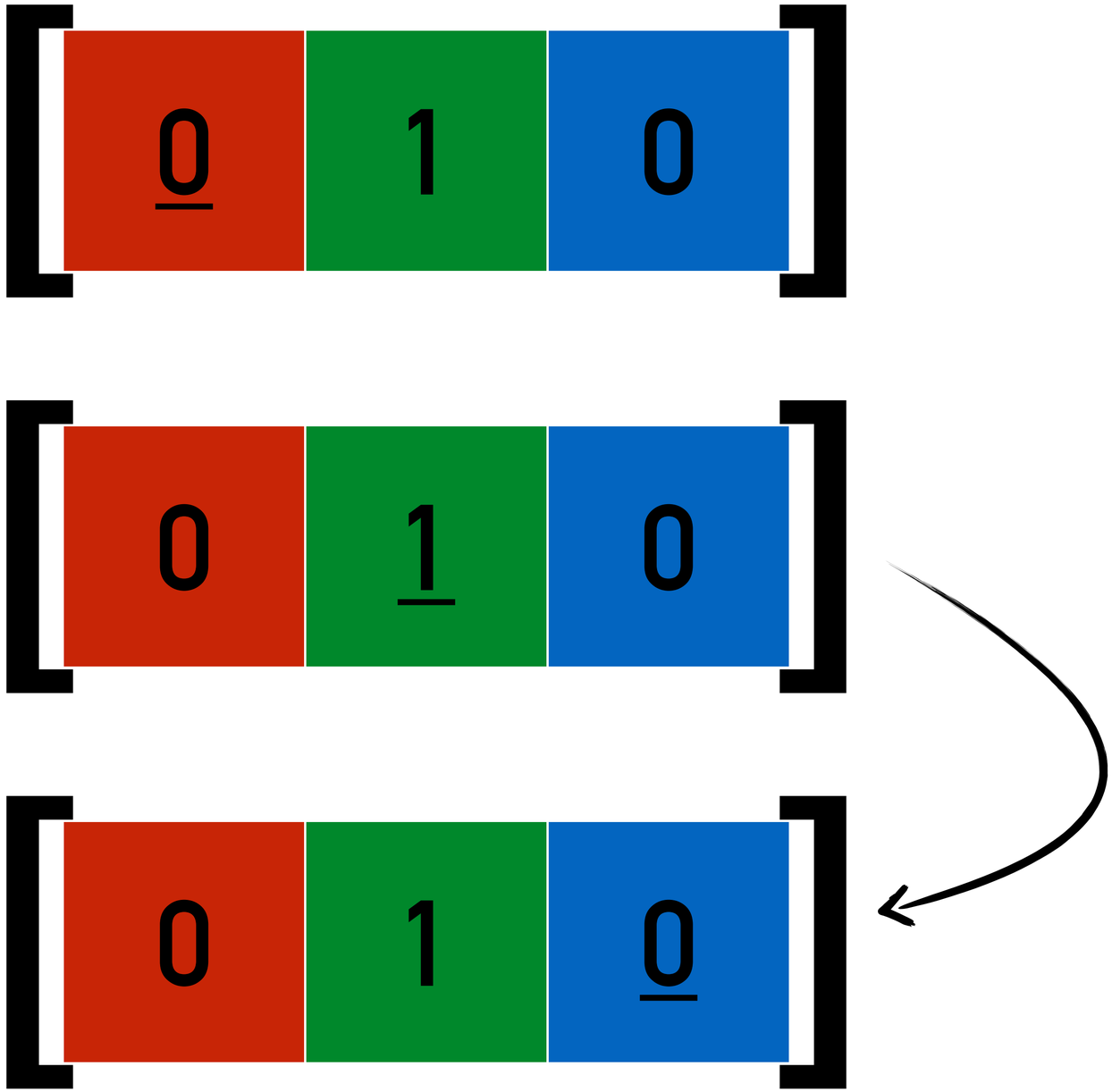
$$1 + 1 = 1 \quad ✘$$

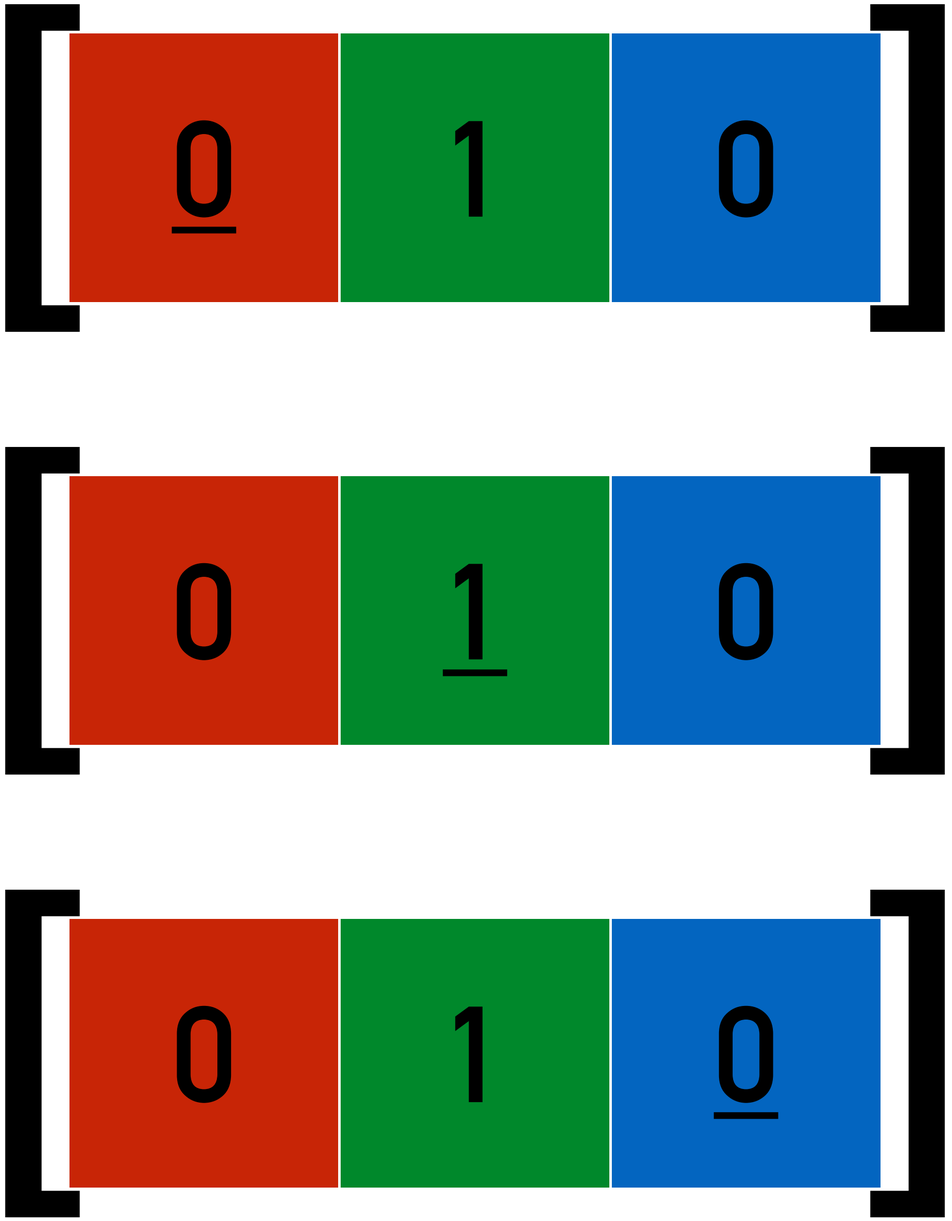

0 0 0
0 0 0
0 0 0

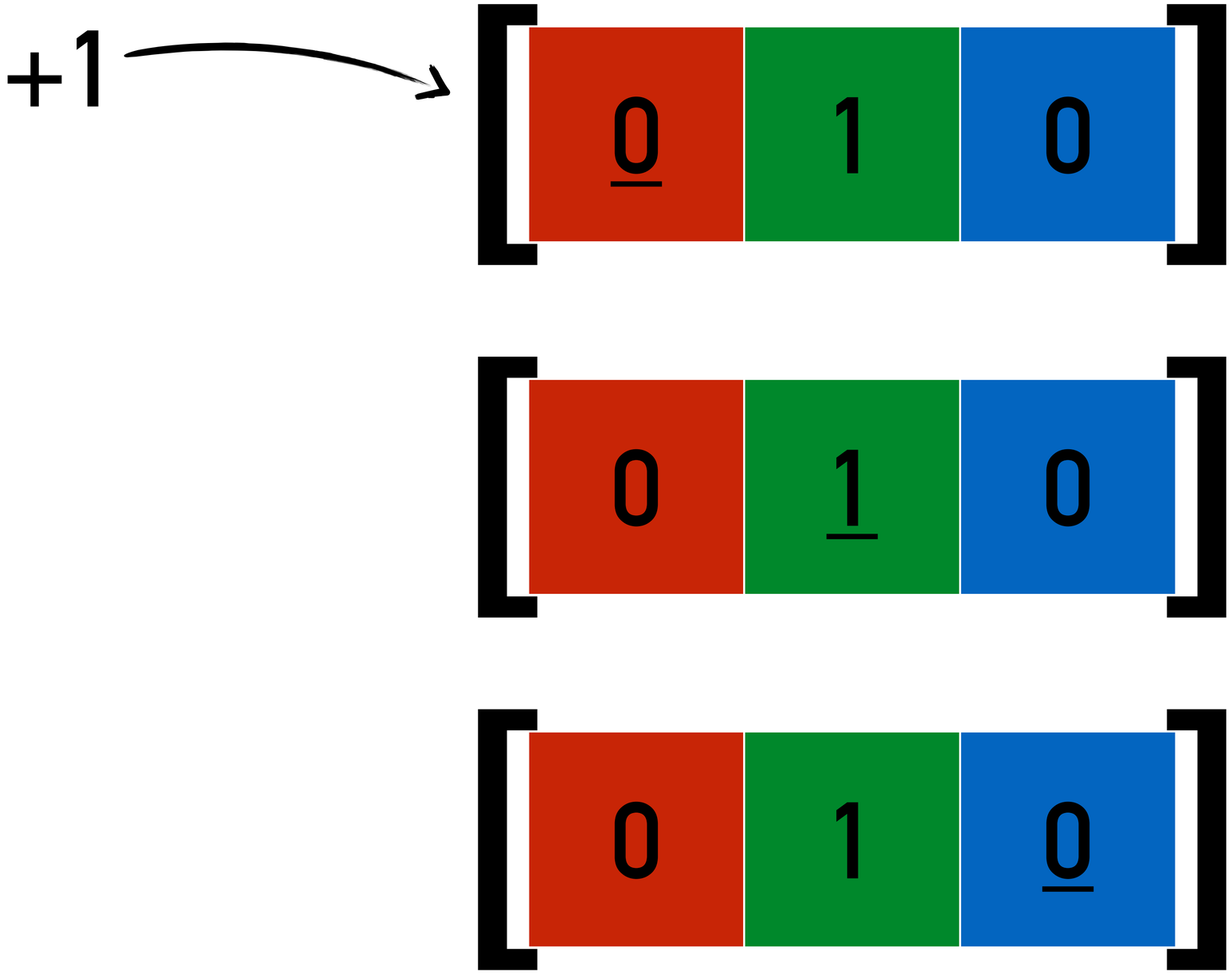

+1
0 0 0
0 0 0
0 0 0

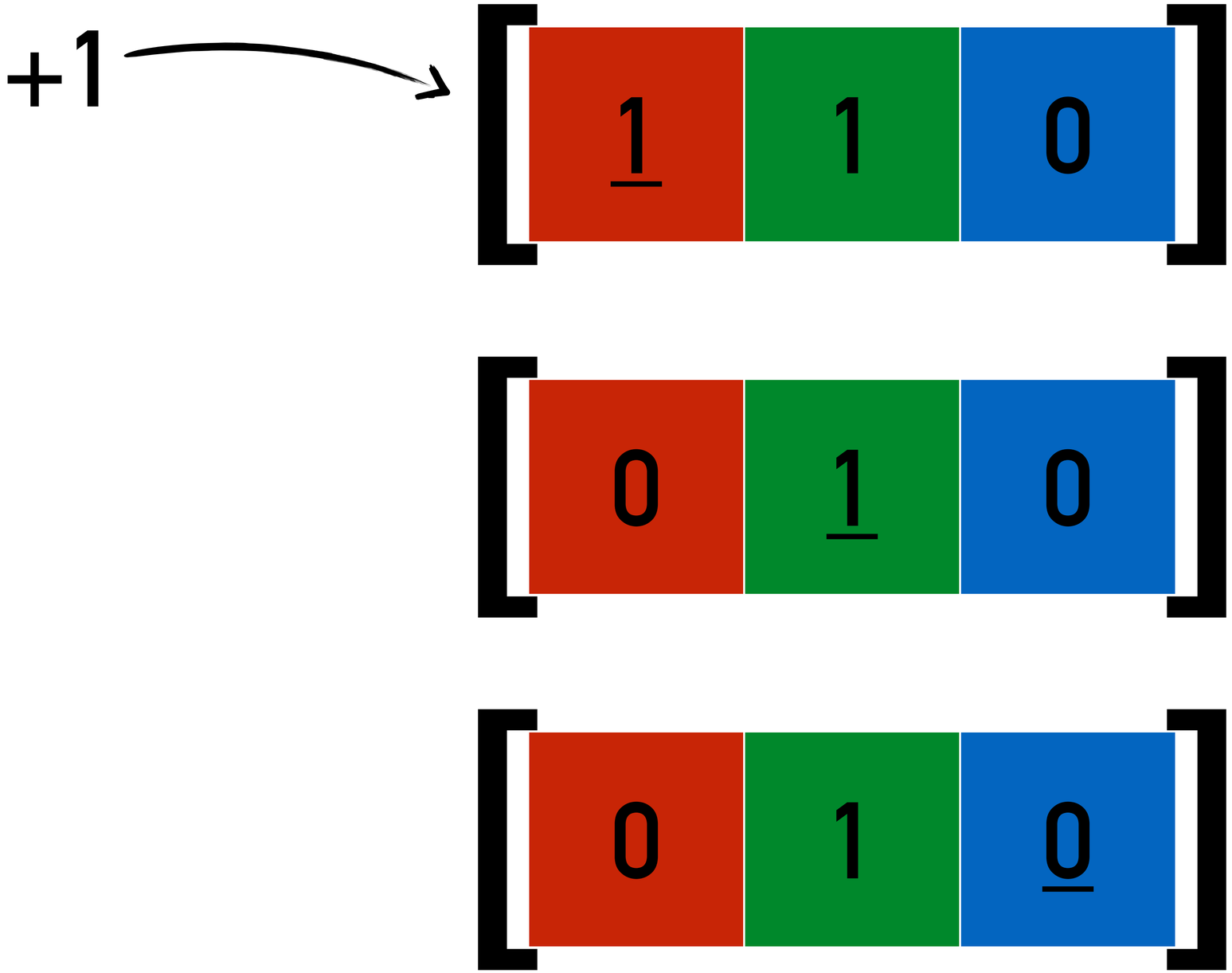

+1
0 0 0
0 1 0
0 0 0

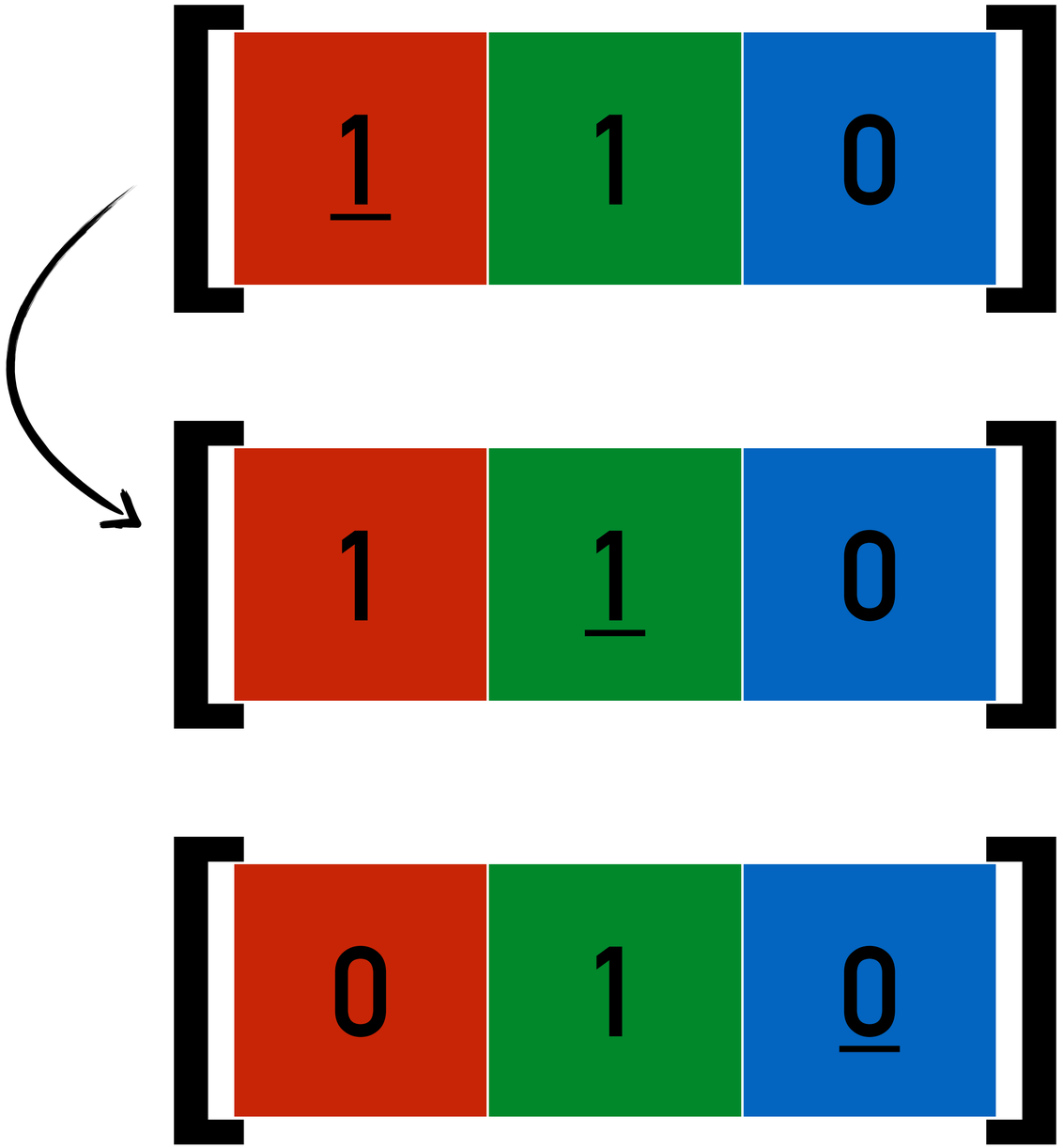

0 1 0
0 1 0
0 0 0

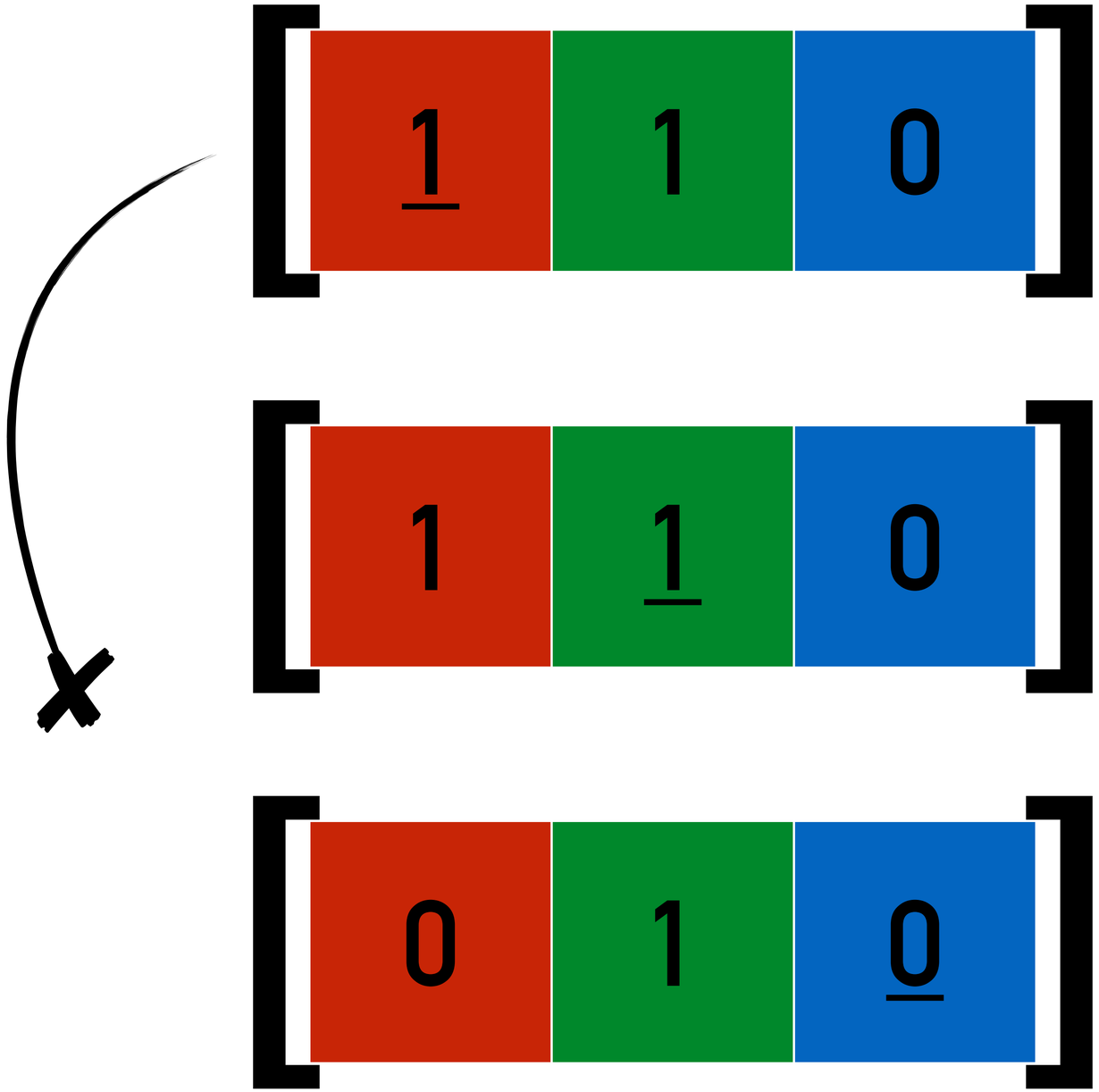

0 1 0
0 1 0
0 1 0

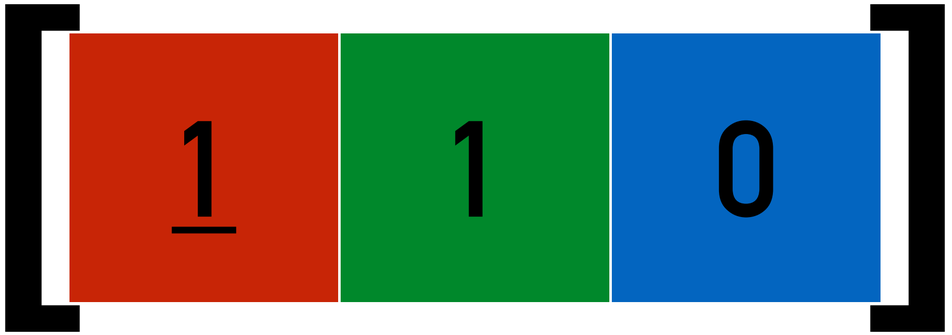

0 1 0
0 1 0
0 1 0

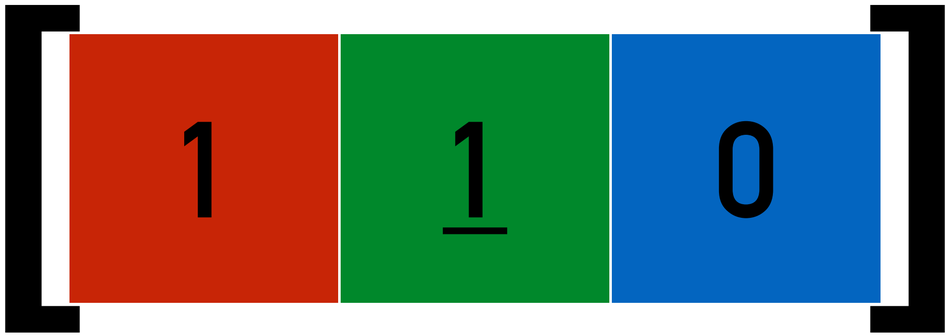

+1
0 1 0
0 1 0
0 1 0

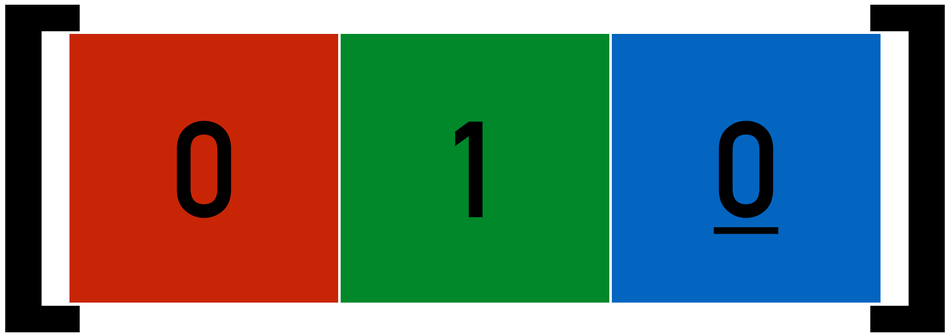

+1
1 1 0
0 1 0
0 1 0

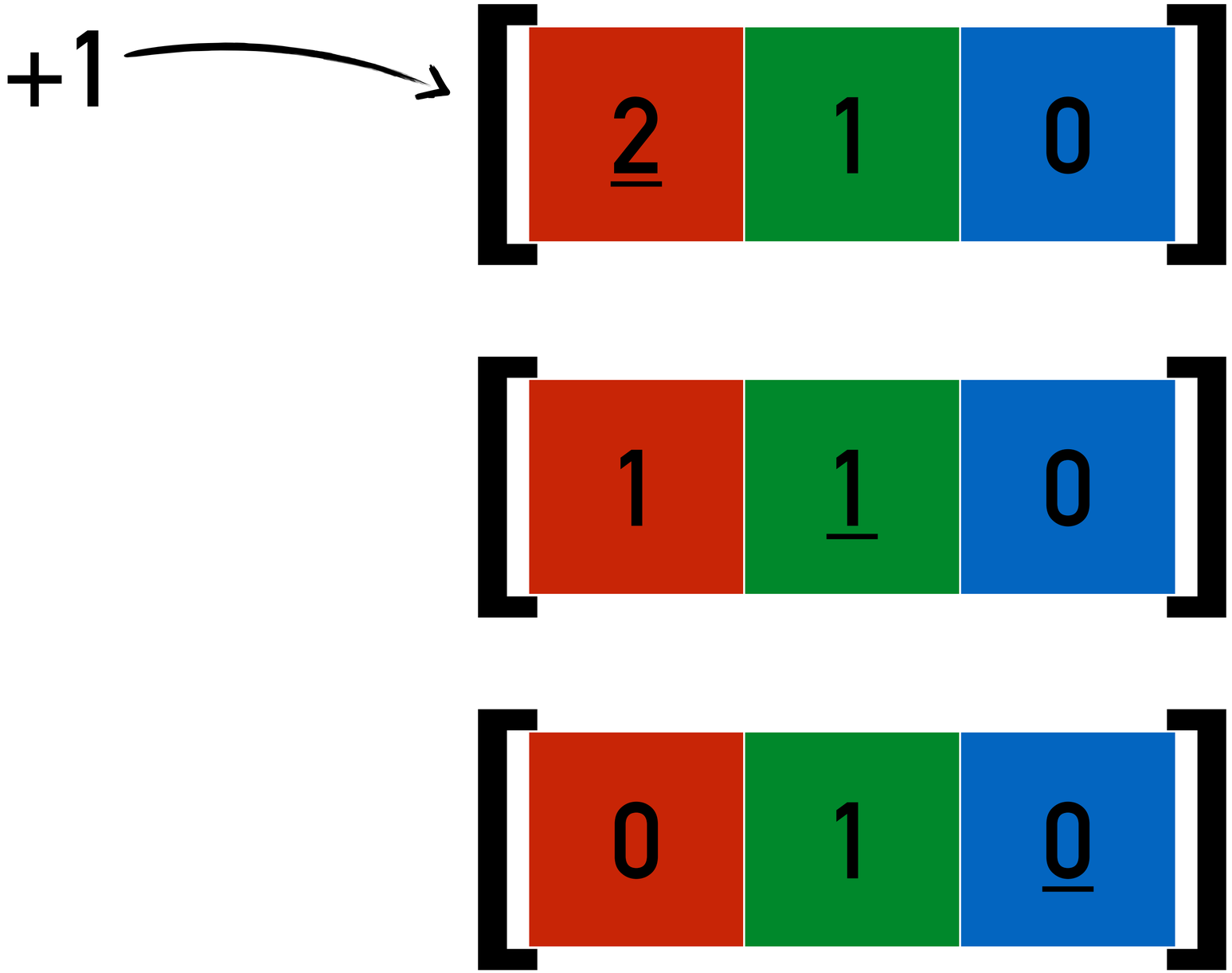

1 1 0
1 1 0
0 1 0

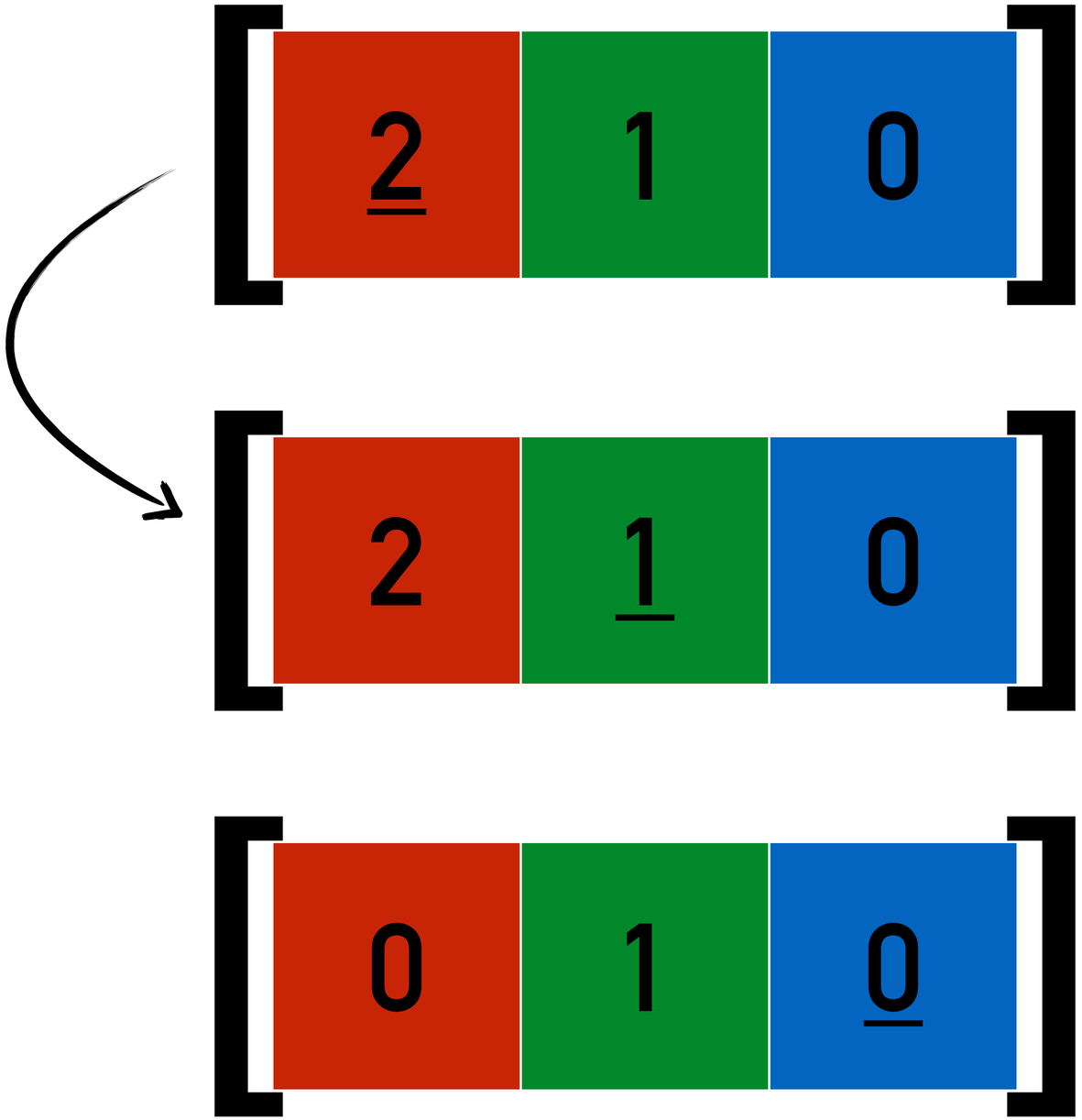

1 1 0
1 1 0
0 1 0

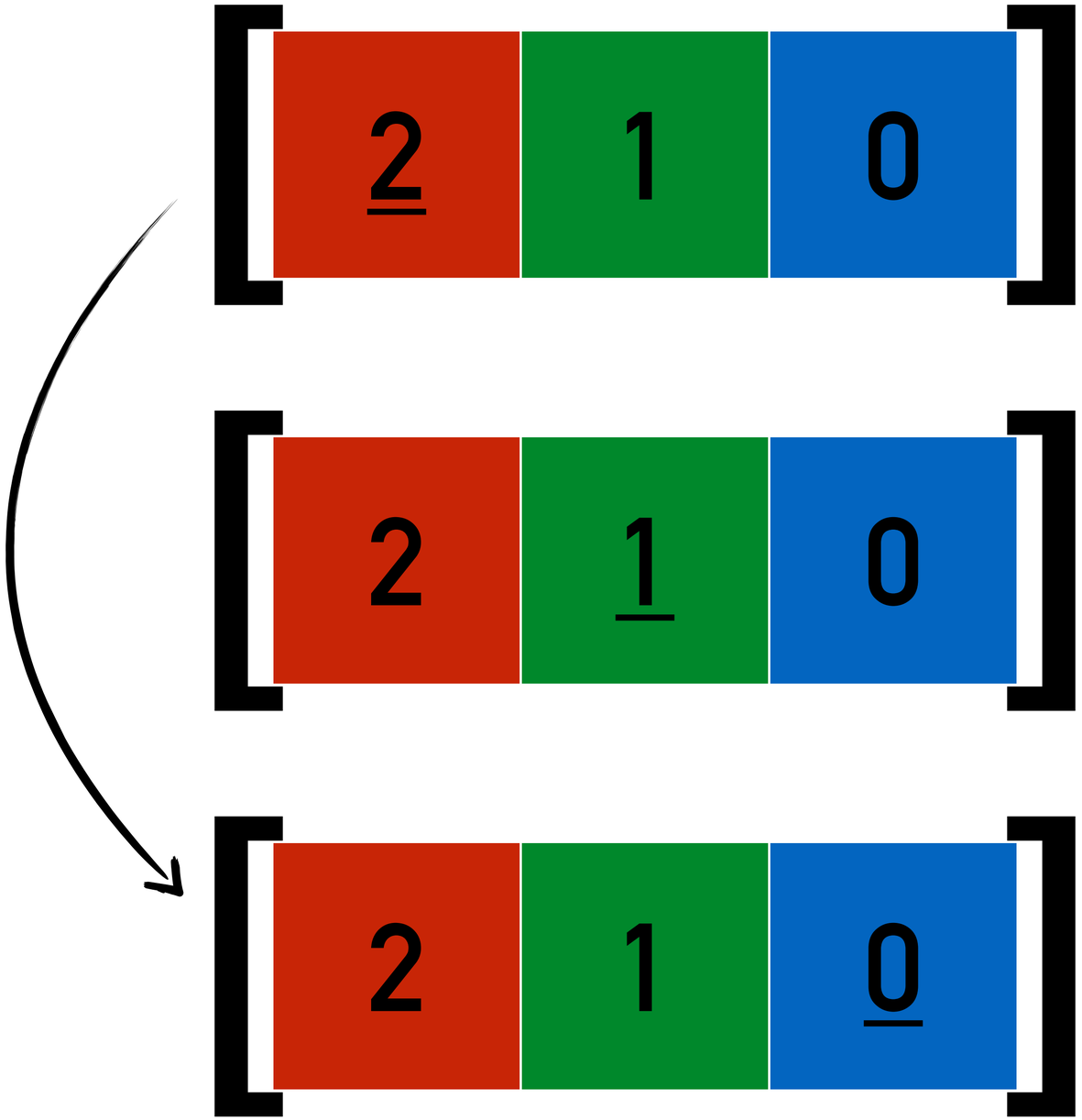

1
1
0

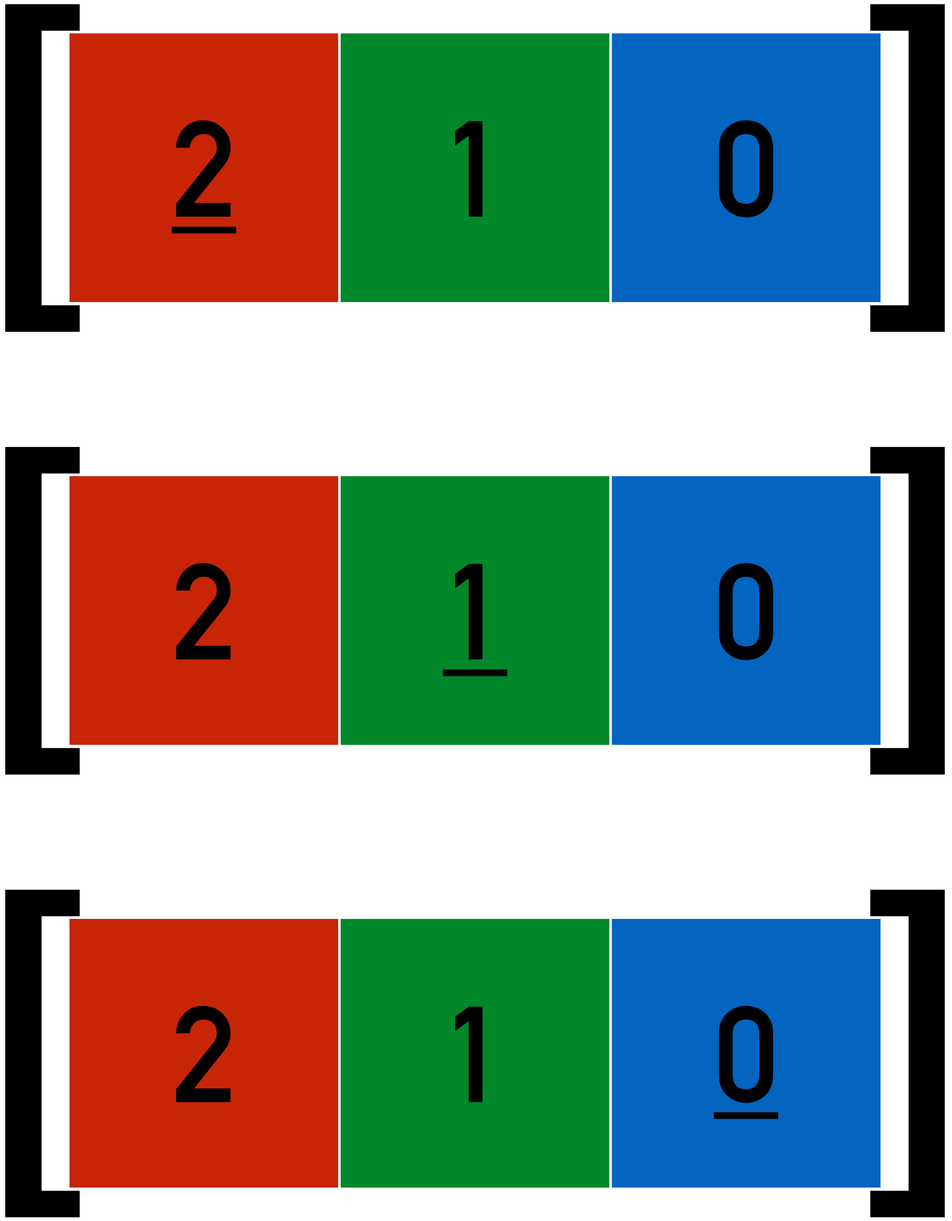

1
1
0

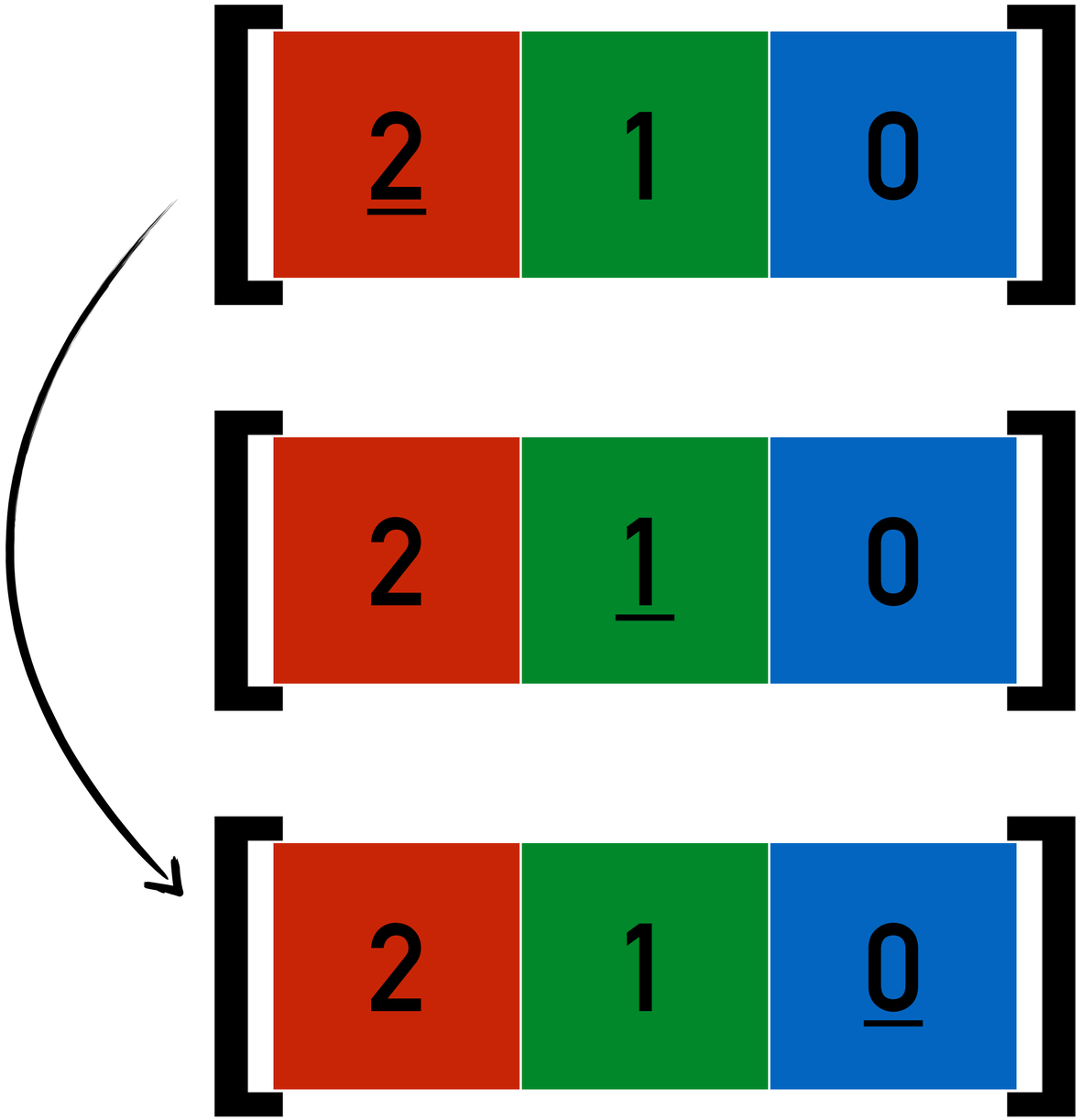

0
1
0

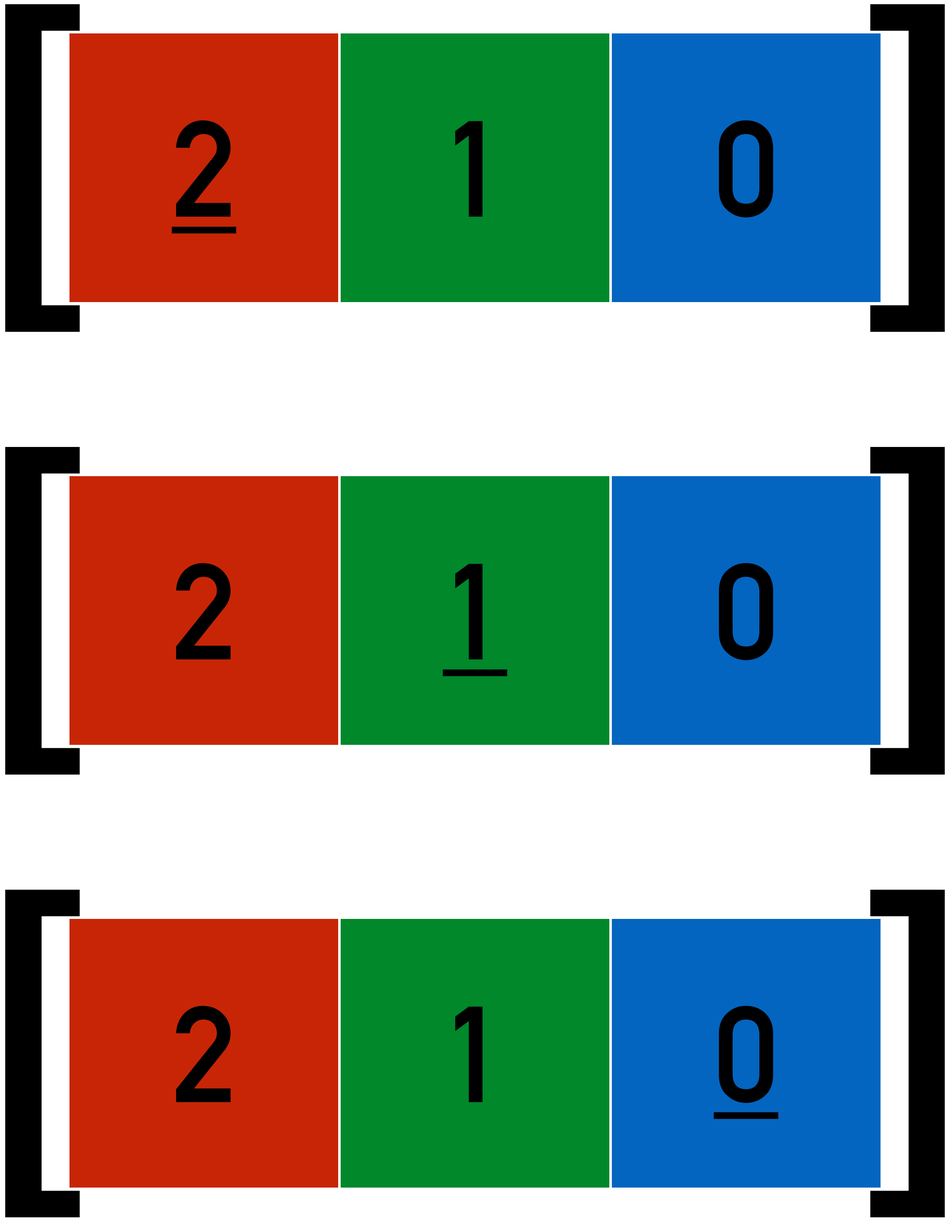

+1
2 1 0
1 1 0
0 1 0


2 1 0
2 1 0
0 1 0


2 1 0
2 1 0
2 1 0


2 1 0
2 1 0
2 1 0


2 1 0
2 1 0
2 1 0


2 1 0
2 1 0
2 1 0

Better increment-only counter

∪

$$\{1\} \cup (\{2\} \cup \{3\}) = (\{1\} \cup \{2\}) \cup \{3\} \quad ✔$$

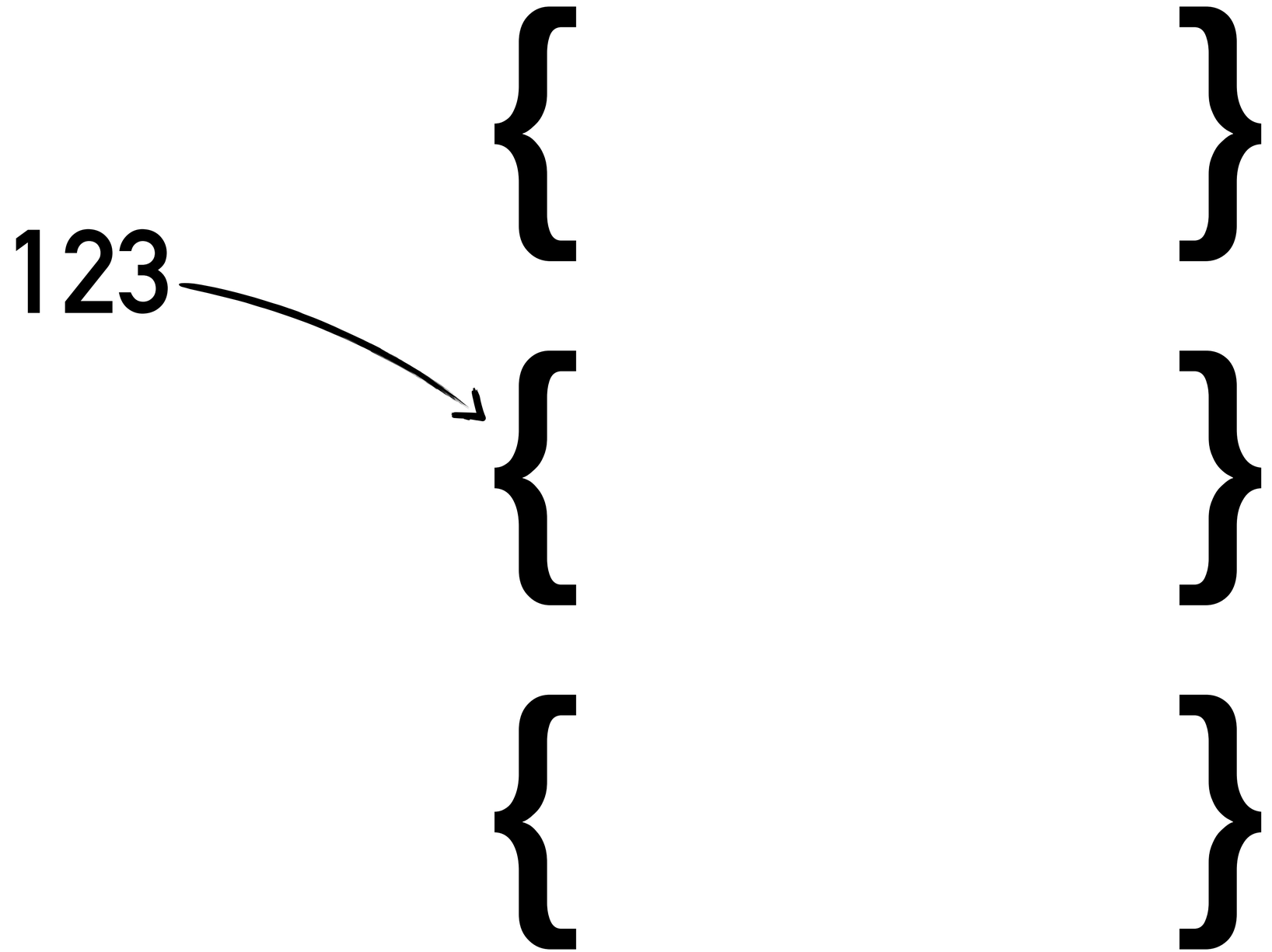
$$\{1\} \cup \{2\} = \{2\} \cup \{1\} \quad ✔$$

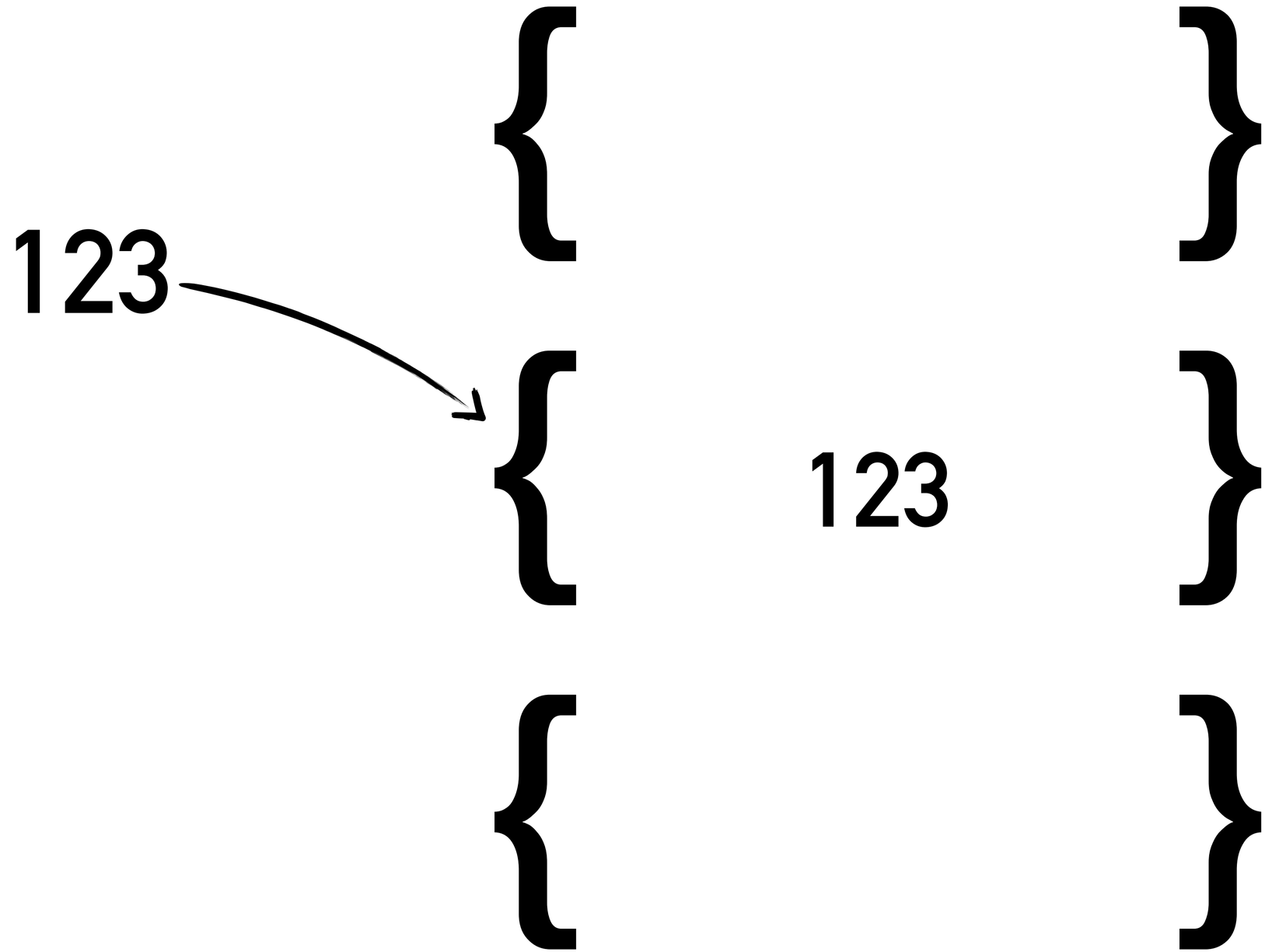
$$\{1\} \cup \{1\} = \{1\} \quad ✔$$

{ }

{ }

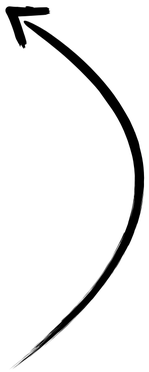
{ }


123
{
}
{
}
{
}


123
{
}
{
123
}
{
}

{ 123 }

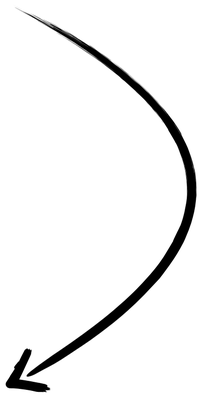
{ 123 }



{ }

{ 123 }

{ 123 }

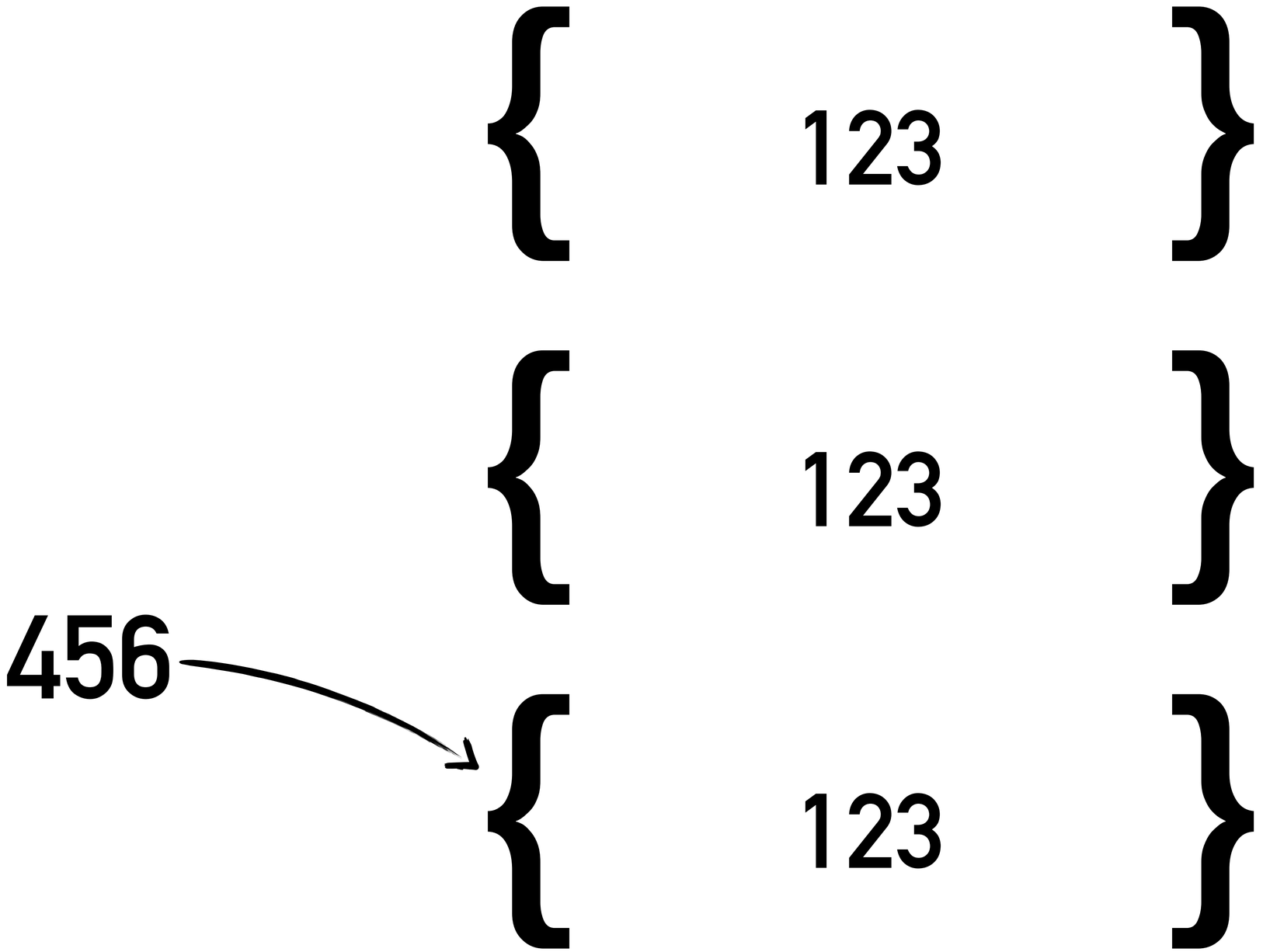


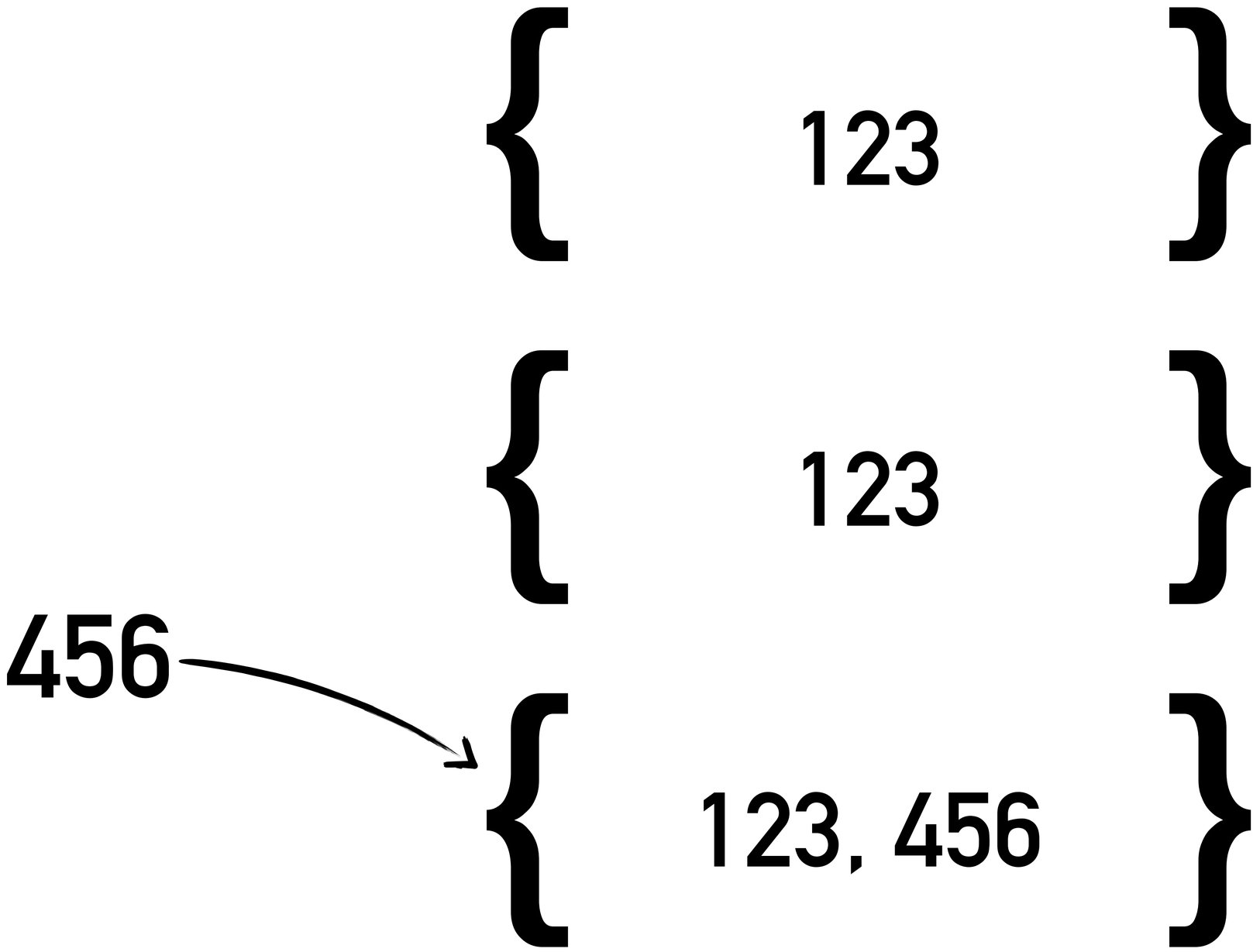
{ 123 }

{ 123 }

{ 123 }

{ 123 }

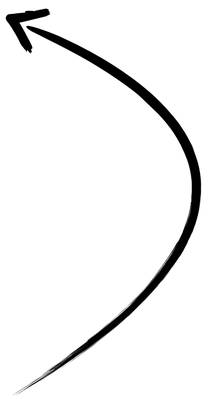

{ 123 }
{ 123 }
456
{ 123 }

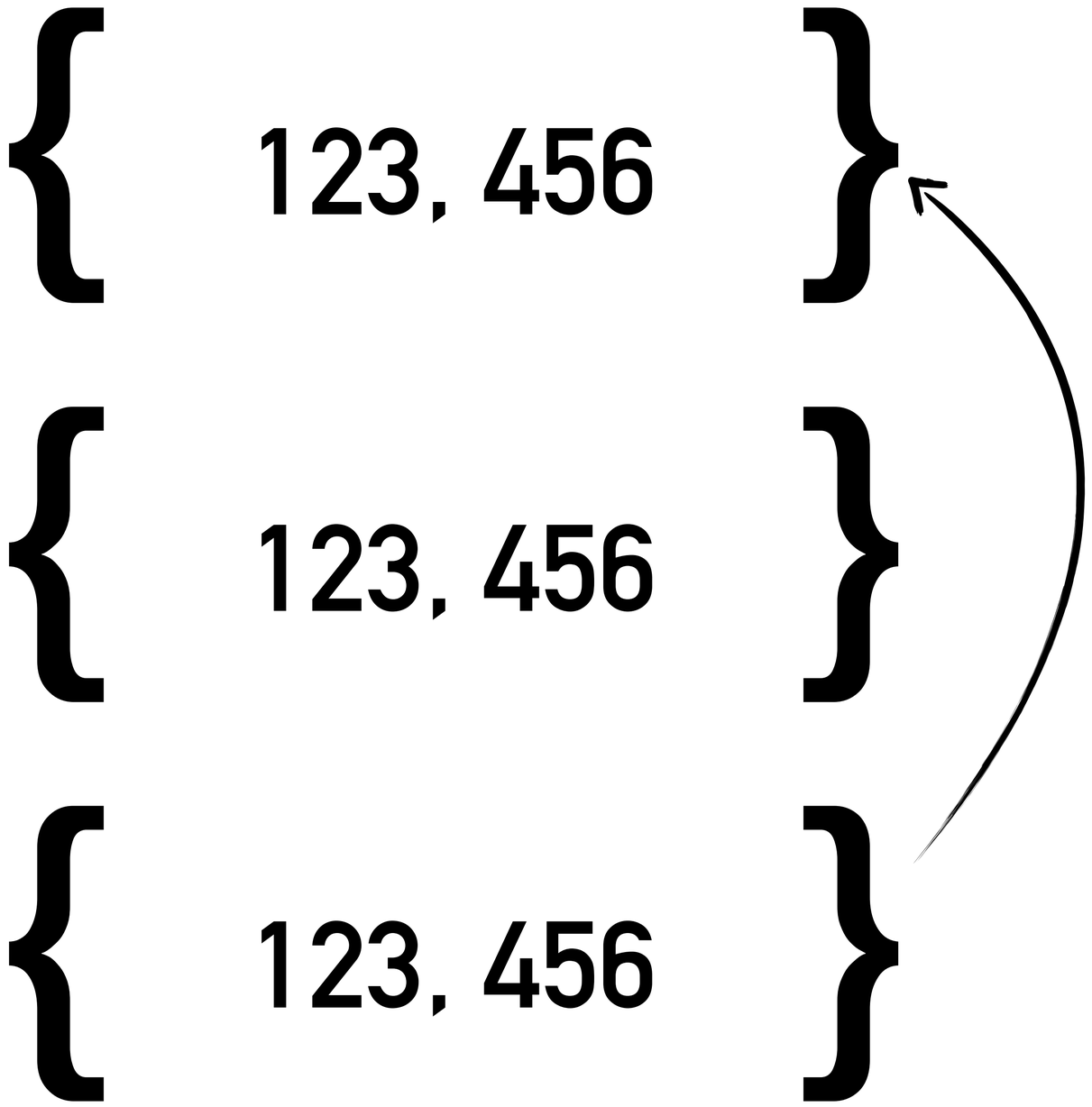

{ 123 }
{ 123 }
456
{ 123, 456 }

{ 123 }

{ 123, 456 }

{ 123, 456 }




{ 123, 456 }
{ 123, 456 }
{ 123, 456 }

{ 123, 456 }

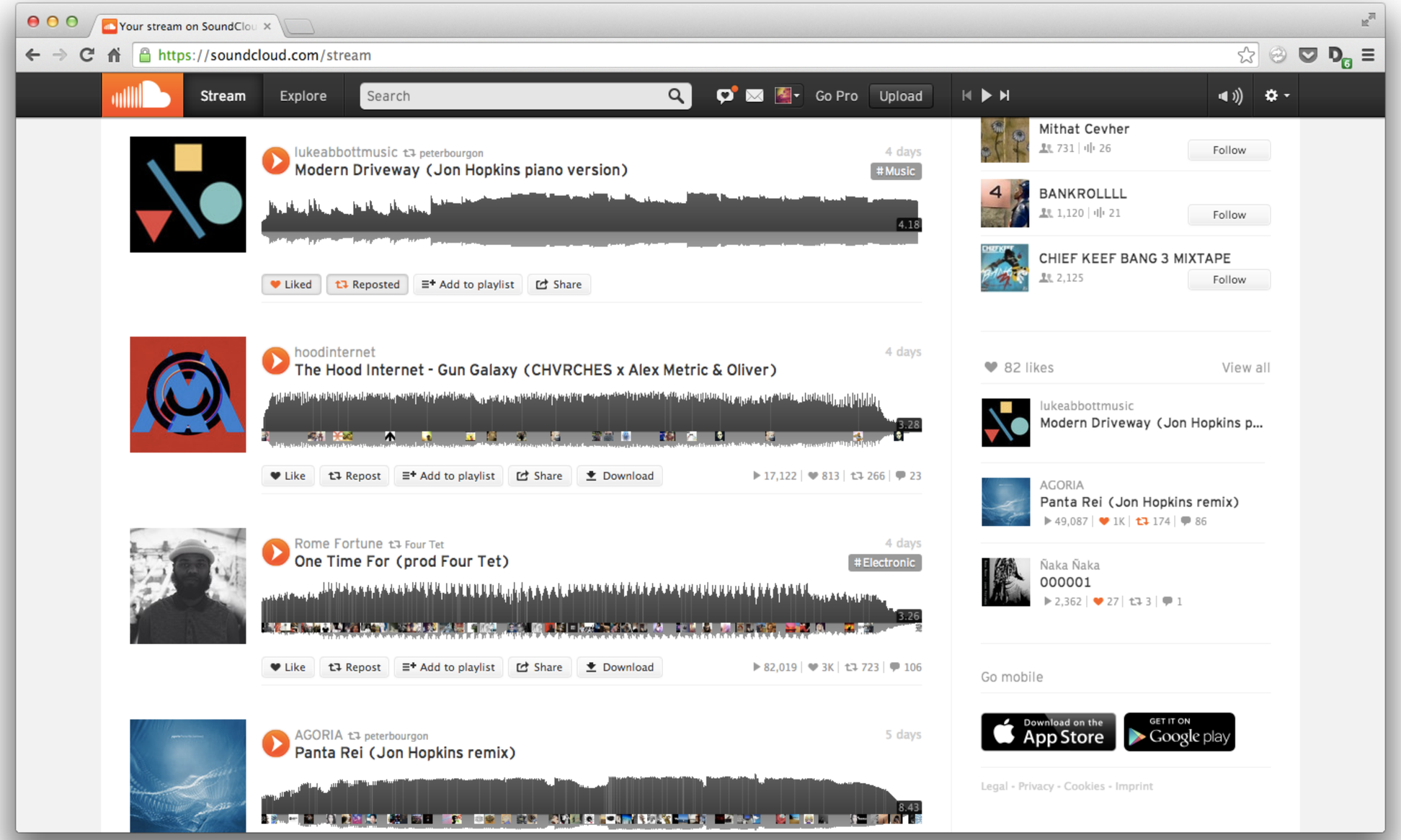
{ 123, 456 }

{ 123, 456 }

Interlude:
Bending the problem

CRDTs in production


SOUNDCLOUD


Your stream on SoundClou
https://soundcloud.com/stream
Stream
Explore
Search
Go Pro
Upload
lukeabbottmusic peterbourgon
Modern Driveway (Jon Hopkins piano version)
4 days
#Music
4.18
Liked
Reposted
Add to playlist
Share
hoodinternet
The Hood Internet - Gun Galaxy (CHVRCHES x Alex Metric & Oliver)
4 days
3.28
Like
Repost
Add to playlist
Share
Download
17,122 | 813 | 266 | 23
Rome Fortune Four Tet
One Time For (prod Four Tet)
4 days
#Electronic
3.26
Like
Repost
Add to playlist
Share
Download
82,019 | 3K | 723 | 106
AGORIA peterbourgon
Panta Rei (Jon Hopkins remix)
5 days
8.43
Mithat Cevher
731 | 26
Follow
BANKROLLLL
1,120 | 21
Follow
CHIEF KEEF BANG 3 MIXTAPE
2,125
Follow
82 likes
View all
lukeabbottmusic
Modern Driveway (Jon Hopkins p...
AGORIA
Panta Rei (Jon Hopkins remix)
49,087 | 1K | 174 | 86
Ñaka Ñaka
000001
2,362 | 27 | 3 | 1
Go mobile
Download on the App Store
GET IT ON Google play
Legal · Privacy · Cookies · Imprint

Event

Timestamp

User

Verb

Identifier

Event

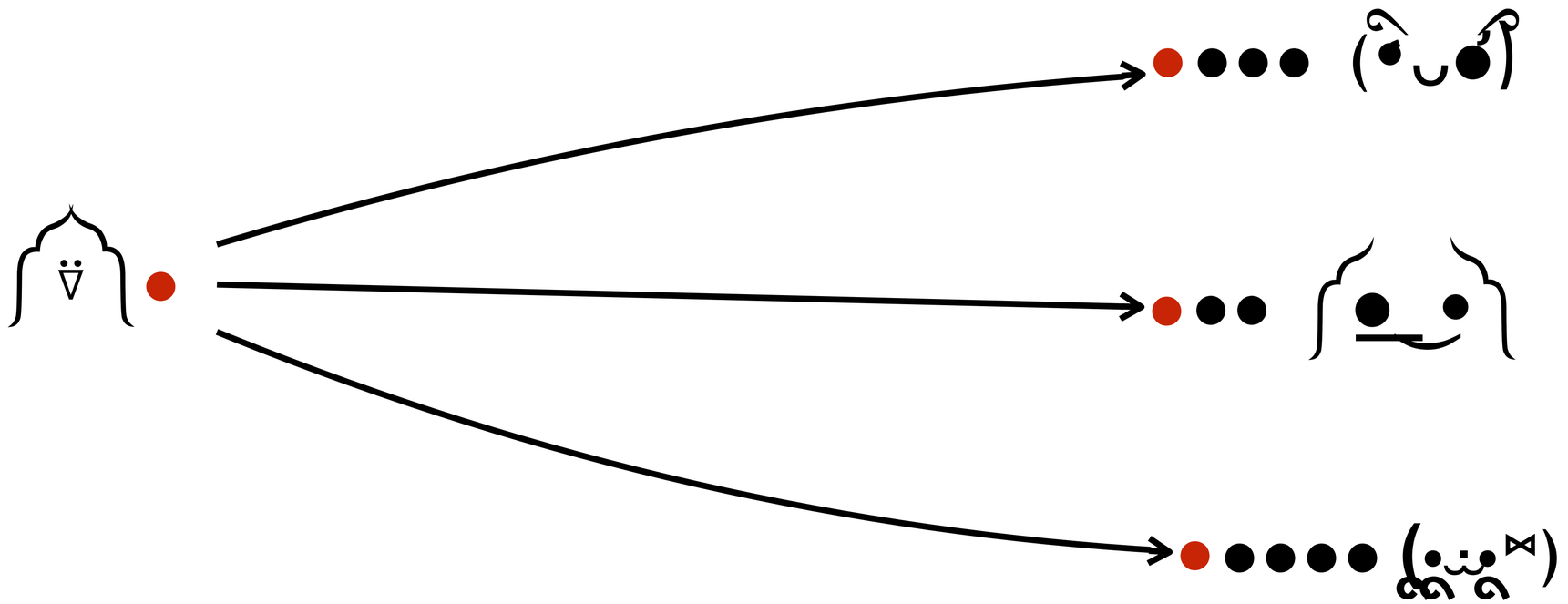
At 2014-05-26 12:04:56.097403 UTC

snoopdogg

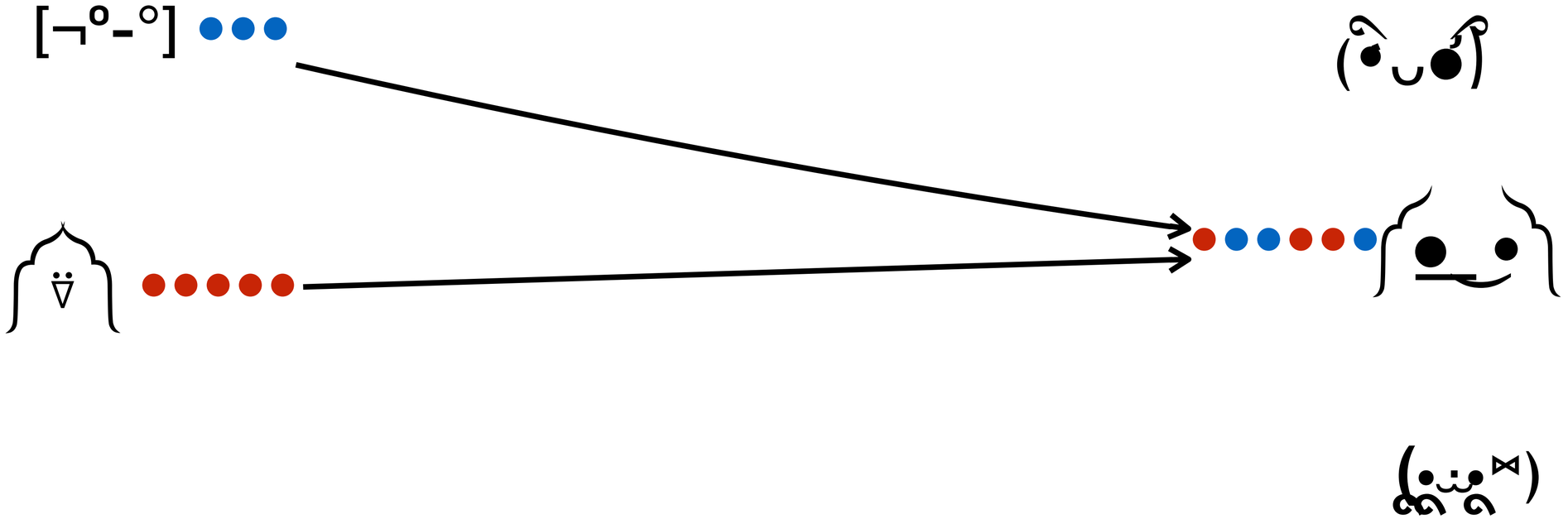
Reposted

theeconomist/election-day

Fan out on write



Fan in on read



Events are unique: use a set

G-set: can’t delete

2P-set: add, remove once

OR-set: storage overhead

CRDT sets

$S+ = \{A\ B\ C\}$

$S- = \{B\}$

$\mathbf{S} = \{A\ C\}$

A new set appears!

New set

S+ = { A/1 B/2 C/3 }

S– = { D/4 }

S = { A/1 B/2 C/3 }

New set

S = actor's set key
e.g. **snoopdogg:outbox**

A, B, C, D = actor:verb:identifier
e.g. **snoop:repost:theeconomist/election-day**

1, 2, 3 = timestamp
e.g. **2014-05-26T12:04:56.097403Z**

Reading
is easy

Writing
is interesting

Insert (key, element, score)

- If either *key*+ or *key*– already contains *element*, and the existing score >= *score*,
  no-op and exit.
- Insert (*element, score*) into add set *key*+.
- Delete (*element*) from remove set *key*–.

Delete (key, element, score)

- If either *key*+ or *key*– already contains *element*, and the existing score >= *score*,
  no-op and exit.
- Insert (*element, score*) into remove set *key*–.
- Delete (*element*) from add set *key*+.

Example

S+ { A/1 B/2 }

S– { C/3 }

Insert D/4

S+ { A/1 B/2 }

S– { C/3 }

Insert D/4

S+ { A/1 B/2 D/4 }

S– { C/3 }

S+ { A/1 B/2 D/4 }

S– { C/3 }

Insert D/4

S+ { A/1 B/2 D/4 }

S– { C/3 }

Insert D/4

S+ { A/1 B/2 D/4 }

S– { C/3 }

S+ { A/1 B/2 D/4 }

S– { C/3 }

Insert D/3

S+ { A/1 B/2 D/4 }

S– { C/3 }

Insert D/3

S+ { A/1 B/2 D/4 }

S– { C/3 }

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/3

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/3

S+ { A/1 B/2 D/4 }

S– { C/3 }

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/4

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/4

S+ { A/1 B/2 D/4 }

S– { C/3 }

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/5

S+ { A/1 B/2 D/4 }

S– { C/3 }

Delete D/5

S+ { A/1 B/2 ~~D/4~~ }

S– { C/3 D/5 }

S+ { A/1 B/2 }

S– { C/3 D/5 }

Insert D/5

S+ { A/1 B/2 }

S– { C/3 D/5 }

Insert D/5

S+ { A/1 B/2 }

S– { C/3 D/5 }

S+ { A/1 B/2 }

S– { C/3 D/5 }

Delete D/6

S+ { A/1 B/2 }

S– { C/3 D/5 }

Delete D/6

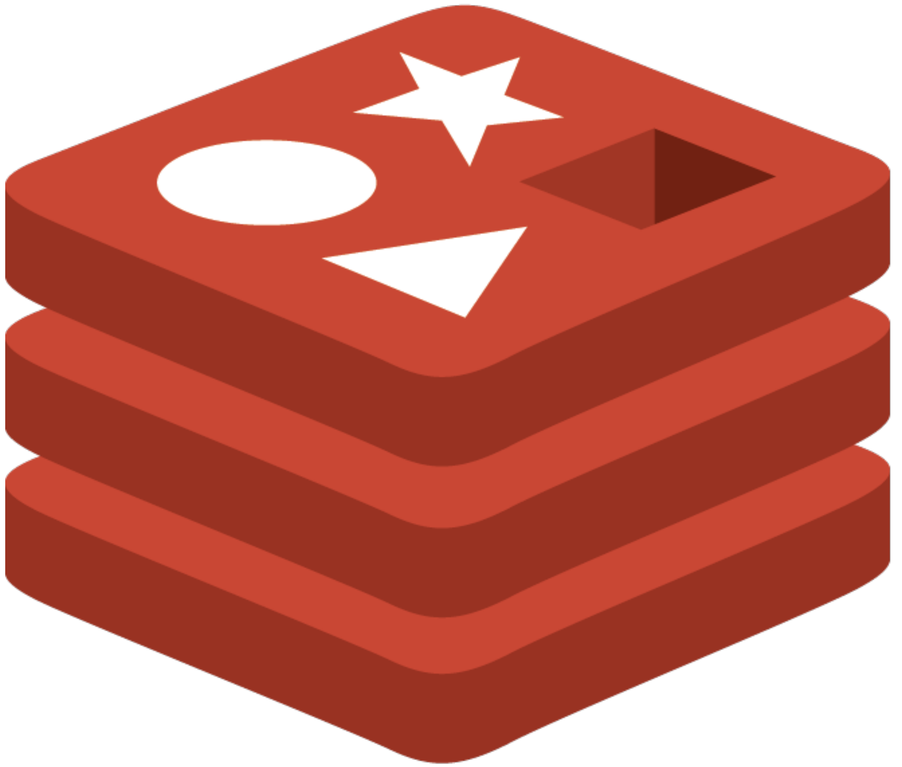
S+ { A/1 B/2 }

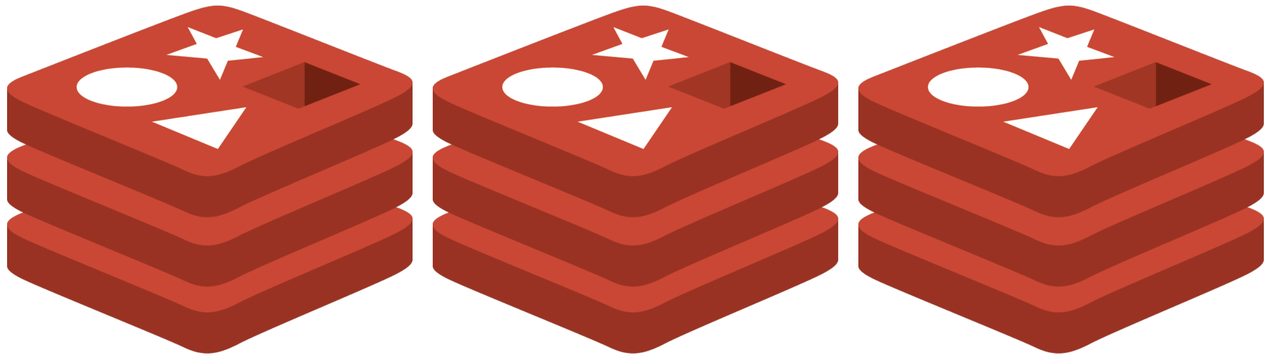
S– { C/3 D/6 }

S+ { A/1 B/2 }

S– { C/3 D/6 }

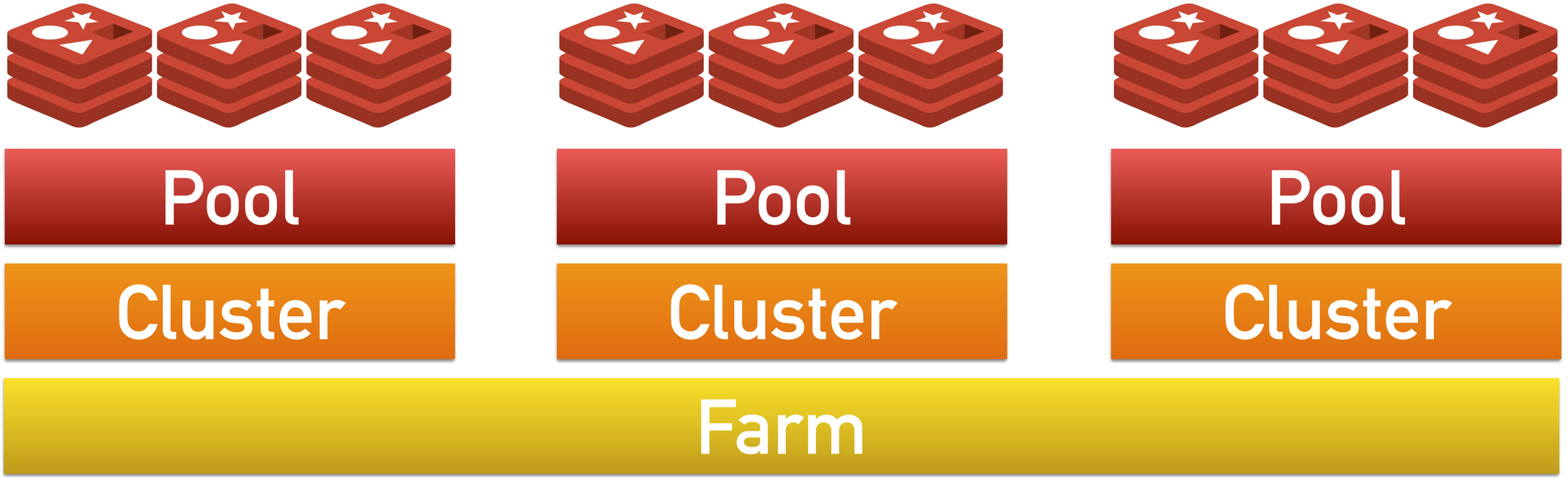
Making it real





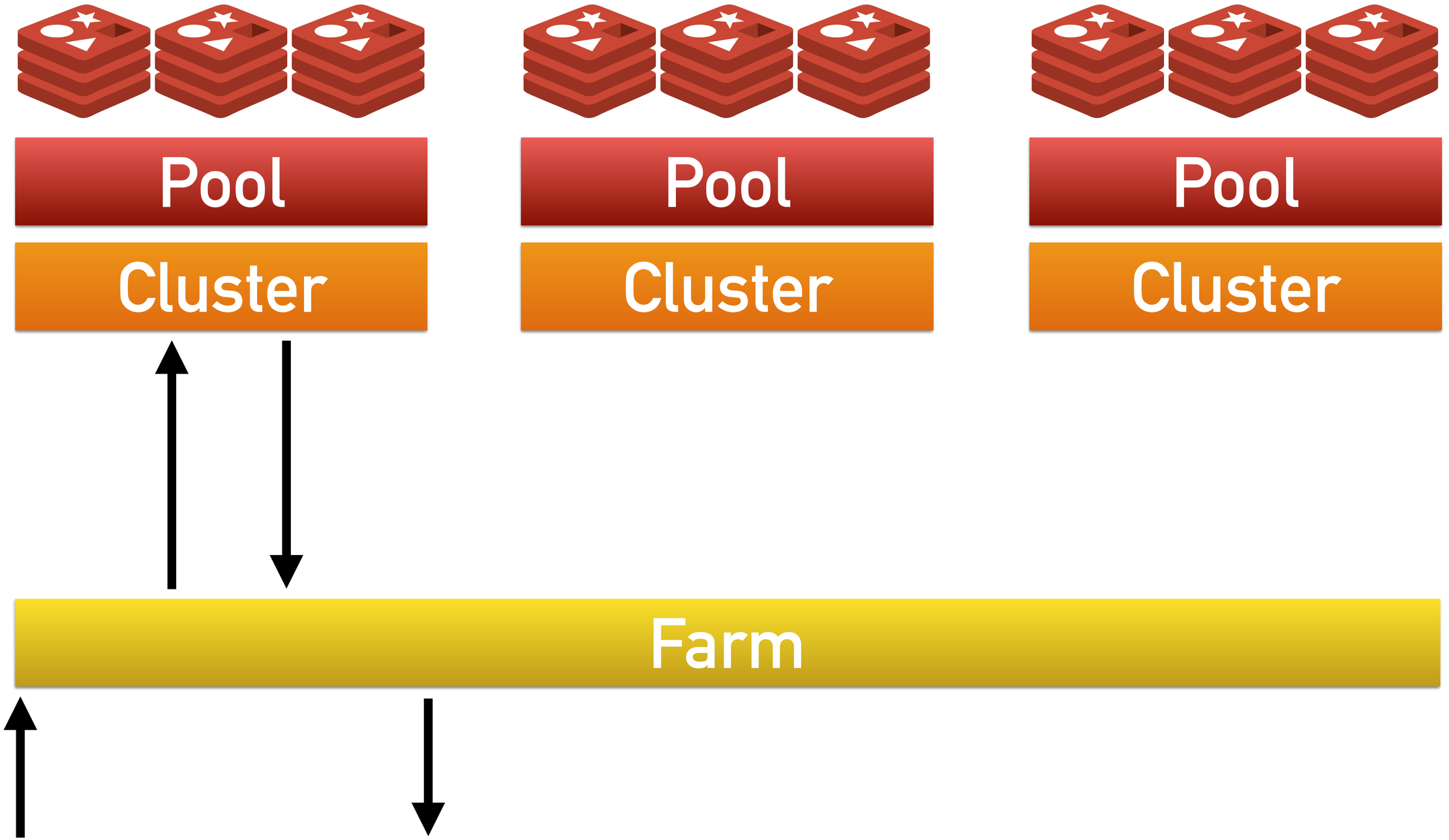
Pool

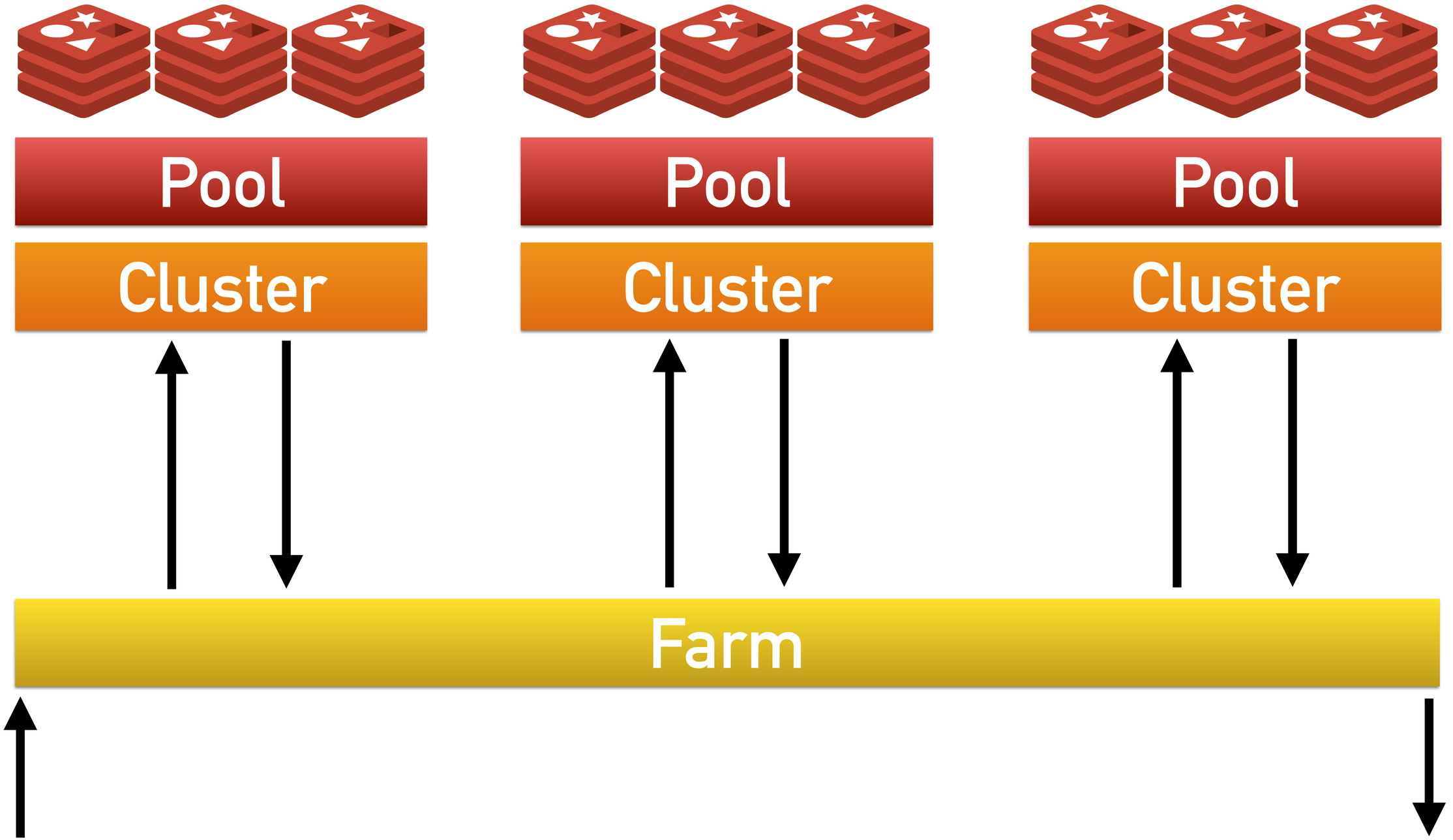
Cluster

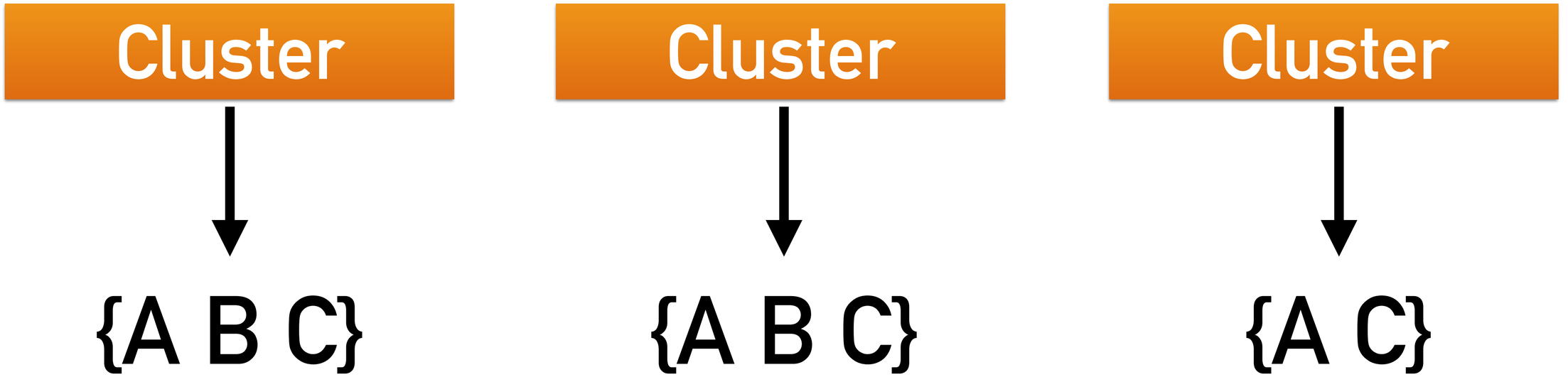

Pool
Pool
Pool
Cluster
Cluster
Cluster
Farm

Writing
is easy

Reading
is interesting

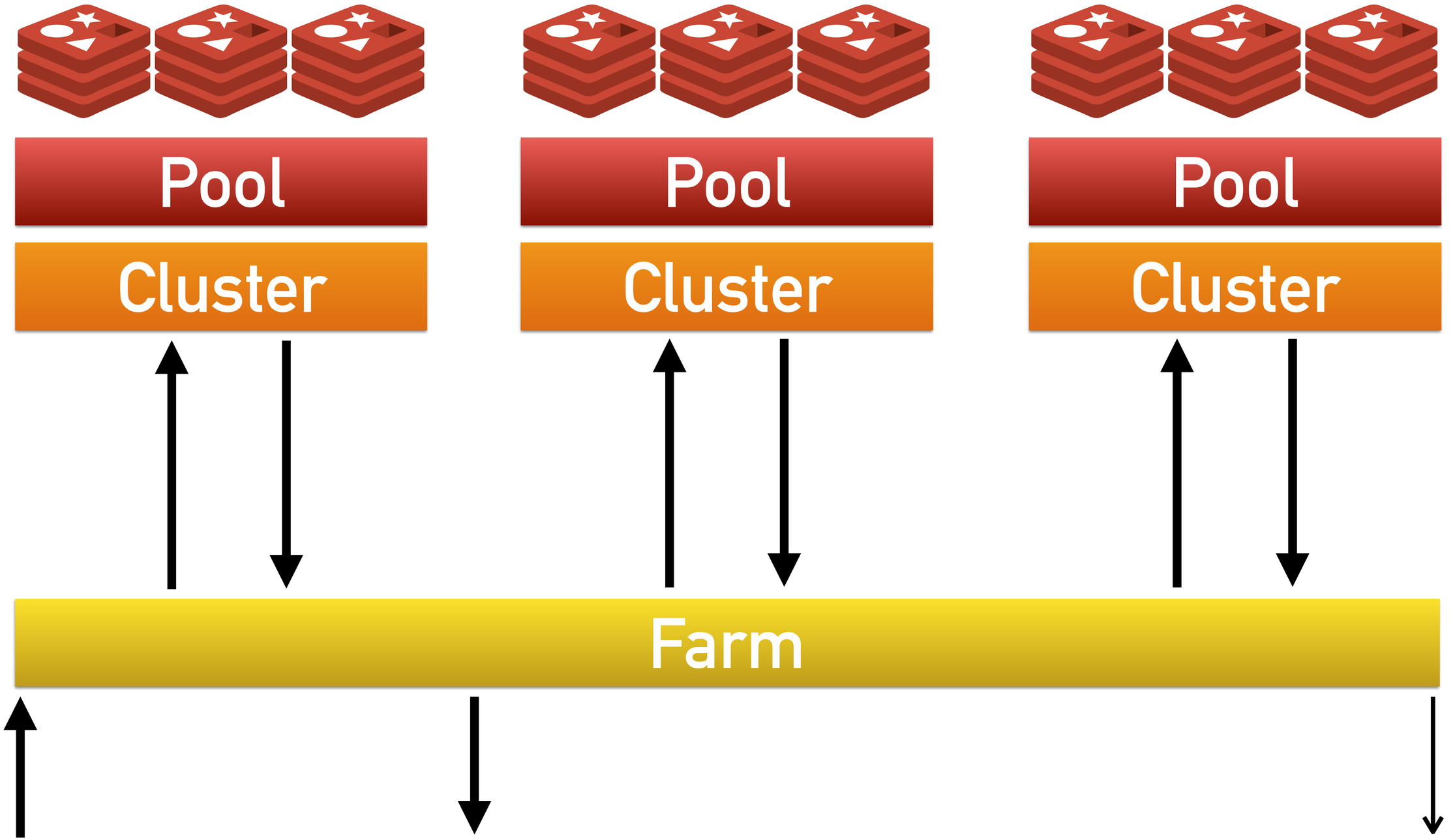

Pool
Pool
Pool
Cluster
Cluster
Cluster
Farm


Pool
Pool
Pool
Cluster
Cluster
Cluster
Farm



$$\cup = \{A\ B\ C\}$$

$$\Delta = \{B\}$$


Pool
Pool
Pool
Cluster
Cluster
Cluster
Farm

github.com/soundcloud/roshi

Conclusions

Consistency without consensus = CRDT.

Embrace your invariants.

Don’t bend a good solution to fit your problem;
bend your problem to fit a good solution.

Thanks a ton!

@peterbourgon

SoundCloud

☞ soundcloud.com/jobs ☜